UỶ BAN NHÂN DÂN
THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN

BÁO CÁO ĐỒ ÁN MÔN HỌC
PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH CẢM XÚC

DỰA TRÊN ĐÁNH GIÁ CỦA SINH VIÊN

Thành viên tham gia:	Trần Vĩnh Huy	- 3122411072
	Nguyễn Minh Nhựt	- 3122411144
	Trần Gia Bảo	- 3122411013
	Trương Quang Long	- 3122411114

Giảng viên hướng dẫn: TS. Trịnh Tấn Đạt
TS. Nguyễn Thị Tuyết Nam

Thành phố Hồ Chí Minh, 05/2025

LỜI CẢM ƠN

Lời đầu tiên, nhóm xin được gửi lời cảm ơn chân thành nhất đến thầy TS. Trịnh Tấn Đạt và cô TS. Nguyễn Thị Tuyết Nam. Trong quá trình học tập và tìm hiểu môn "Phân tích dữ liệu", nhóm đã nhận được rất nhiều sự quan tâm, giúp đỡ, hướng dẫn tâm huyết và tận tình của thầy và cô. Thầy cô đã giúp em tích lũy thêm nhiều kiến thức về môn học này để có thể hoàn thành được bài báo cáo đồ án môn học về đề tài: *"Phân tích cảm xúc dựa trên đánh giá của sinh viên"*

Trong quá trình làm bài chắc chắn khó tránh khỏi những thiếu sót. Do đó, nhóm kính mong nhận được những lời góp ý của thầy để bài tiểu luận của em ngày càng hoàn thiện hơn.

Nhóm xin chân thành cảm ơn!

PHẦN MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh giáo dục Việt Nam không ngừng đổi mới theo hướng lấy người học làm trung tâm, việc lắng nghe và phân tích phản hồi từ sinh viên đóng vai trò ngày càng quan trọng trong việc nâng cao chất lượng giảng dạy và quản lý đào tạo.

Tuy nhiên, phần lớn các đánh giá sinh viên hiện nay vẫn được xử lý thủ công hoặc chỉ dựa trên thống kê định lượng, chưa khai thác hiệu quả nguồn dữ liệu văn bản tự do phong phú chứa đựng nhiều sắc thái cảm xúc và thông tin ngữ nghĩa sâu sắc.

Trước thực tế đó, đề tài *"Phân tích cảm xúc dựa trên đánh giá của sinh viên"* được lựa chọn nhằm xây dựng một hệ thống phân tích cảm xúc tự động trên văn bản phản hồi của sinh viên, từ đó cung cấp một công cụ hỗ trợ nhà trường phát hiện các vấn đề tiềm ẩn trong hoạt động giảng dạy, điều chỉnh phương pháp giảng dạy kịp thời, cũng như ghi nhận những điểm mạnh đáng khuyến khích. Việc áp dụng các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên (NLP) hiện đại không chỉ góp phần số hóa công tác khảo sát và đánh giá mà còn mang lại hướng tiếp cận mới, khách quan và hiệu quả hơn trong việc cải tiến chất lượng giáo dục đại học tại Việt Nam.

2. Mục đích và nhiệm vụ nghiên cứu

Mục đích của nghiên cứu này là xây dựng và đánh giá hiệu quả của các mô hình học máy trong việc phân tích cảm xúc từ văn bản phản hồi của sinh viên nhằm hỗ trợ các cơ sở giáo dục nâng cao chất lượng giảng dạy, cải thiện dịch vụ đào tạo, và thúc đẩy sự tương tác tích cực giữa giảng viên và người học. Qua đó, nghiên cứu hướng đến việc phát triển một công cụ ứng dụng có khả năng tự động nhận diện cảm xúc tích cực, tiêu cực hoặc trung lập từ ý kiến sinh viên, góp phần vào tiến trình chuyển đổi số trong quản lý giáo dục đại học tại Việt Nam.

Để đạt được mục đích trên, nghiên cứu thực hiện các nhiệm vụ cụ thể sau:

1. Khảo sát và tổng quan các phương pháp phân tích cảm xúc trên văn bản tiếng Việt, đặc biệt trong bối cảnh giáo dục.

2. Thu thập và tiền xử lý tập dữ liệu đánh giá của sinh viên để đảm bảo tính sạch và phù hợp cho huấn luyện mô hình.
3. Xây dựng và triển mô hình phân tích cảm xúc
4. So sánh hiệu suất các mô hình trên cùng tập dữ liệu theo các chỉ số đánh giá chuẩn như Accuracy, F1-score,…
5. Đề xuất mô hình tối ưu và thảo luận tiềm năng ứng dụng vào thực tiễn phân tích cảm xúc trong môi trường giáo dục.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các phương pháp phân tích cảm xúc trên văn bản tiếng Việt, đặc biệt là các kỹ thuật xử lý ngôn ngữ tự nhiên ứng dụng trong việc phân loại cảm xúc từ nội dung phản hồi của sinh viên đối với môn học, giảng viên hoặc chương trình đào tạo. Bên cạnh đó, đề tài cũng tập trung khảo sát và áp dụng các mô hình học máy (machine learning) và học sâu (deep learning) như TF-IDF kết hợp SVM, FastText embedding, và mô hình ngữ cảnh PhoBERT nhằm đánh giá hiệu quả xử lý văn bản trong ngữ cảnh tiếng Việt.

Phạm vi của đề tài giới hạn trong việc xử lý và phân tích các phản hồi dạng văn bản thu thập được từ sinh viên tại một số cơ sở giáo dục đại học tại Việt Nam. Các phản hồi được gán nhãn cảm xúc theo ba mức: tích cực, tiêu cực và trung lập. Nghiên cứu không đi sâu vào các ngôn ngữ khác ngoài tiếng Việt và không mở rộng sang các loại dữ liệu phi văn bản (như âm thanh, hình ảnh). Các mô hình được đánh giá dựa trên dữ liệu có sẵn, không triển khai thành hệ thống hoàn chỉnh nhưng có tiềm năng làm nền tảng cho các ứng dụng thực tiễn trong quản lý và cải tiến giáo dục.

4. Phương pháp nghiên cứu

Đề tài sử dụng phương pháp nghiên cứu thực nghiệm định lượng kết hợp với các kỹ thuật trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và học máy (machine learning) để xây dựng và đánh giá hiệu quả của các mô hình phân tích cảm xúc. Quá trình nghiên cứu được triển khai qua các bước chính như sau:

1. Khảo sát lý thuyết và thu thập dữ liệu: Trước tiên, nghiên cứu tiến hành

khảo sát các công trình liên quan đến phân tích cảm xúc, đặc biệt trong bối cảnh tiếng Việt. Đồng thời, một tập dữ liệu đánh giá môn học từ sinh viên được thu thập từ hệ thống khảo sát nội bộ tại một số trường đại học.

2. Tiền xử lý dữ liệu văn bản: Dữ liệu thu thập được tiến hành các bước xử lý cơ bản như chuẩn hóa văn bản (xóa ký tự đặc biệt, chuẩn hóa chữ viết tắt, viết thường,...), tách từ, loại bỏ stop words và chuyển đổi văn bản thành dạng biểu diễn số để phục vụ huấn luyện mô hình. Mỗi hướng tiếp cận sẽ sử dụng kỹ thuật tiền xử lý phù hợp với đặc trưng của phương pháp biểu diễn văn bản.
3. Biểu diễn đặc trưng văn bản theo:
    a. Hướng truyền thống: sử dụng Bag-of-Words với TF-IDF từ bộ từ điển của UIT-VSFC.
    b. Hướng embedding: sử dụng FastText để chuyển văn bản thành các vector nhúng dựa trên ký tự n-gram.
    c. Hướng ngữ cảnh: sử dụng PhoBERT để trích xuất vector ngữ nghĩa sâu dựa trên kiến trúc Transformer.
4. Huấn luyện và đánh giá mô hình phân loại: Các biểu diễn văn bản được đưa vào huấn luyện bằng nhiều thuật toán học máy khác nhau: Random Forest, Support Vector Machine, XGBoost, và Neural Network. Mỗi mô hình được đánh giá trên cùng một tập dữ liệu thử nghiệm bằng các chỉ số như Accuracy, Precision, Recall và F1-score để đưa ra so sánh hiệu quả giữa các hướng tiếp cận.
5. Phân tích kết quả và đề xuất mô hình tối ưu: Kết quả thu được từ các mô hình sẽ được phân tích định lượng và định tính để đánh giá hiệu suất từng hướng tiếp cận. Dựa trên các chỉ số đánh giá, nghiên cứu đề xuất mô hình tối ưu cho bài toán phân tích cảm xúc từ văn bản đánh giá sinh viên, đồng thời thảo luận những ưu và nhược điểm của từng hướng.

5. Những đóng góp mới của đề tài

Đề tài "Phân tích cảm xúc dựa trên đánh giá của sinh viên" mang lại một số đóng góp học thuật và thực tiễn đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt và ứng dụng vào giáo dục đại học, cụ thể như sau:

1. Xây dựng một quy trình chuẩn để tiền xử lý và biểu diễn văn bản tiếng Việt trong bối cảnh giáo dục.
2. So sánh hệ thống giữa ba hướng tiếp cận mô hình biểu diễn và phân loại cảm xúc.
3. Đề xuất mô hình hiệu quả nhất cho bài toán phân tích cảm xúc trong giáo dục đại học Việt Nam.
4. Tiềm năng mở rộng và ứng dụng trong các hệ thống hỗ trợ ra quyết định

Kết quả nghiên cứu đặt nền tảng cho việc phát triển các hệ thống tự động phân tích phản hồi người học, phục vụ quản lý chất lượng giảng dạy, cải tiến chương trình học, hoặc thậm chí phát hiện sớm các vấn đề trong trải nghiệm học tập của sinh viên.

6. Cấu trúc của đề tài

Ngoài phần mở đầu, kết luận, danh mục tài liệu tham khảo và phụ lục, đề tài gồm có 5 chương:

Chương 1: Giới thiệu vấn đề nghiên cứu

Chương 2: Cơ sở lý thuyết

Chương 3: Mô hình đề xuất

Chương 4: Thực nghiệm và đánh giá kết quả

MỤC LỤC

DANH MỤC BẢNG BIỂU, HÌNH ẢNH

DANH MỤC CHỮ VIẾT TẮT

TF	Term Frequency
IDF	Inverse Document Frequency
NLP	Neuro-linguistic programming
RF	Random Forest
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting
NN	Neural Network
BoW	Bag of Words
CBOW	Continuous Bag of Words
BERT	Bidirectional Encoder Representations from Transformers

BẢN TÓM TẮT ĐỀ TÀI

Đề tài *"Phân tích cảm xúc dựa trên đánh giá của sinh viên"* được thực hiện nhằm đề xuất một hệ thống có khả năng tự động phân loại cảm xúc trong văn bản đánh giá nhằm nâng cao chất lượng giảng dạy và trải nghiệm học tập của sinh viên từ việc khai thác thông tin từ các đánh giá môn học của sinh viên

Nghiên cứu triển khai ba hướng tiếp cận biểu diễn văn bản bao gồm: (1) phương pháp truyền thống sử dụng Bag-of-Words và TF-IDF từ bộ từ điển UIT-VSFC, (2) biểu diễn từ nhúng FastText dựa trên n-gram ký tự, và (3) mô hình ngôn ngữ hiện đại PhoBERT với khả năng học ngữ cảnh sâu. Các đặc trưng văn bản sau xử lý sẽ được huấn luyện bằng nhiều mô hình phân loại như Random Forest, SVM, XGBoost và mạng nơ-ron. Dữ liệu đầu vào được xây dựng từ tập đánh giá thực tế của sinh viên và được gán nhãn cảm xúc phục vụ cho việc huấn luyện và kiểm thử mô hình.

Kết quả thực nghiệm cho thấy sự khác biệt rõ rệt về hiệu suất giữa các hướng tiếp cận, trong đó mô hình sử dụng PhoBERT kết hợp với SVM hoặc mạng nơ-ron cho kết quả vượt trội về độ chính xác và khả năng tổng quát hóa. Tuy nhiên, nhìn chung kết quả vẫn chưa hoàn hảo; đặc biệt, ranh giới phân biệt giữa hai lớp cảm xúc tích cực và trung lập còn khá mờ nhạt do đặc tính biểu cảm không rõ ràng trong nhiều câu đánh giá. Điều này đặt ra yêu cầu tiếp tục nghiên cứu mở rộng về chiến lược gán nhãn dữ liệu, kết hợp tri thức ngữ nghĩa và khai thác thông tin theo ngữ cảnh rộng hơn nhằm cải thiện độ chính xác trong các nghiên cứu tiếp theo.

PHẦN NỘI DUNG

CHƯƠNG 1: GIỚI THIỆU VẤN ĐỀ NGHIÊN CỨU

1.1. Nhu cầu thực tế

Trong bối cảnh toàn cầu hóa và chuyển đổi số đang diễn ra mạnh mẽ, việc nâng cao chất lượng giáo dục đại học không chỉ là mục tiêu chiến lược của mỗi quốc gia mà còn là yêu cầu tất yếu để đáp ứng nhu cầu xã hội về nguồn nhân lực chất lượng cao. Trên thế giới, các quốc gia phát triển đang liên tục cải tiến chương trình giảng dạy, phương pháp đào tạo cũng như mô hình quản lý đại học nhằm bắt kịp những thay đổi nhanh chóng của tri thức và công nghệ. Tại Việt Nam, vấn đề nâng cao chất lượng giáo dục đại học ngày càng được quan tâm sâu sắc trong bối cảnh hội nhập quốc tế và đổi mới giáo dục toàn diện. Tuy nhiên, thực tiễn vẫn cho thấy khoảng cách đáng kể giữa lý thuyết đào tạo và nhu cầu thực tế của người học, nhà tuyển dụng, cũng như xã hội.

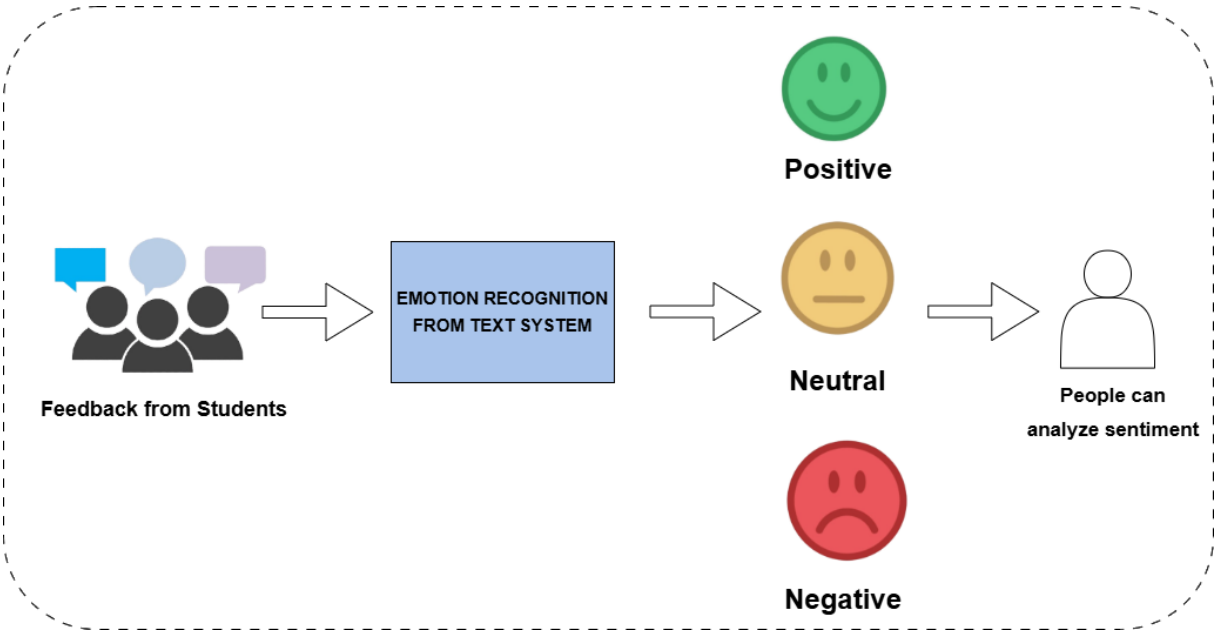
Một trong những yếu tố then chốt góp phần cải thiện chất lượng giáo dục chính là sự tham gia tích cực từ phía người học – cụ thể là sinh viên. Sinh viên không chỉ là đối tượng thụ hưởng dịch vụ giáo dục mà còn là nguồn thông tin phản hồi quan trọng giúp nhà trường điều chỉnh, hoàn thiện hoạt động giảng dạy và quản lý. Những phản hồi này thường được thể hiện dưới dạng các văn bản đánh giá về khóa học, giảng viên, nội dung giảng dạy hoặc trải nghiệm học tập nói chung. Theo chỉ đạo của Thủ tướng Chính phủ, giáo dục và đào tạo phải ***"lấy học sinh làm trung tâm, nhà trường là nền tảng, giáo viên là động lực"*** [1], do đó, việc lắng nghe và phân tích ý kiến đánh giá của sinh viên là cực kỳ quan trọng – không chỉ giúp phát hiện những bất cập trong chương trình đào tạo mà còn thể hiện sự tôn trọng đối với vai trò chủ thể của người học trong quá trình giáo dục. Tuy nhiên, việc xử lý các văn bản đánh giá truyền thống thường tốn thời gian, chủ quan và không tận dụng được hết giá trị dữ liệu tiềm năng từ góc nhìn cảm xúc – yếu tố phản ánh trực tiếp sự hài lòng, bức xúc hoặc kỳ vọng của người học.

Từ thực tiễn đó, một bài toán mới mẻ và có tính ứng dụng cao đã được đặt ra: ***"Phân tích cảm xúc dựa trên đánh giá của sinh viên"***. Đây là một hướng tiếp cận hiện đại trong lĩnh vực xử lý ngôn ngữ tự nhiên NLP và trí tuệ nhân tạo, cho phép tự

động phân tích cảm xúc ẩn sau các phản hồi của sinh viên, từ đó cung cấp cho nhà quản lý giáo dục cái nhìn sâu sắc hơn và khách quan hơn về chất lượng đào tạo. Việc triển khai các hệ thống đánh giá cảm xúc không chỉ giúp rút ngắn thời gian xử lý phản hồi mà còn góp phần nâng cao tính minh bạch, phản biện và cải tiến liên tục trong môi trường giáo dục đại học.

1.2. Bài toán xác định cảm xúc dựa trên văn bản

Hệ thống phân tích cảm xúc là một hệ thống được thiết kế nhằm xác định và phân loại cảm xúc của người viết dựa trên các tín hiệu ngôn ngữ trong văn bản. Hệ thống đánh giá cảm xúc từ văn bản đánh giá của sinh viên là một ứng dụng cụ thể trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), nhằm khai thác và hiểu rõ hơn những phản hồi của người học trong môi trường giáo dục. Kỹ thuật phân tích cảm xúc trong trường hợp này được thực hiện thông qua việc kiểm tra mức độ tương thích giữa văn bản đầu vào với các mẫu cảm xúc đã được gán nhãn sẵn, từ đó đưa ra dự đoán thuộc các lớp cảm xúc như tích cực, tiêu cực hoặc trung tính như hình 1.1.



Hình 1.1. Một hệ thống nhận diện cảm xúc thông qua văn bản

Một hệ thống phân tích cảm xúc từ văn bản thường được xây dựng theo ba hướng tiếp cận chính, phản ánh mức độ phát triển của công nghệ và khả năng xử lý ngôn ngữ tự nhiên ngày càng cao.

1. **Hệ thống phân tích từ khóa** – mô hình cơ bản: là hệ thống sử dụng tập hợp

các từ khóa cảm xúc định nghĩa thủ công (như “hài lòng”, “không thích”, “tệ”, “xuất sắc”) để gán nhãn cảm xúc cho văn bản. Kỹ thuật này hoạt động theo phương pháp so khớp đơn giản một-nhiều giữa văn bản đầu vào và danh sách từ khóa đã định nghĩa. Tuy nhiên, độ chính xác của hệ thống phụ thuộc mạnh vào chất lượng và tính bao quát của tập từ khóa, và không thể xử lý được các cấu trúc câu phức tạp hay ngữ nghĩa ngụ ý.

2. **Hệ thống phân tích học máy truyền thống**: Là hệ thống được thiết kế bằng cách sử dụng một tập văn bản đánh giá đã được gán nhãn cảm xúc để huấn luyện các mô hình học máy như Naive Bayes, SVM hoặc Logistic Regression. Các mô hình này học được các đặc trưng ngôn ngữ (thường là từ vựng và tần suất) để dự đoán cảm xúc văn bản đầu vào. Kỹ thuật phân tích cảm xúc này vẫn sử dụng phương pháp kiểm tra một-nhiều như đã nêu, song độ chính xác được cải thiện đáng kể nhờ quá trình huấn luyện trên dữ liệu thực tế.
3. **Hệ thống phân tích học sâu – ngữ cảnh động**: là hệ thống được phát triển dựa trên các mô hình học sâu như LSTM, BERT hoặc các biến thể của Transformer. Các hệ thống này cho phép khai thác ngữ nghĩa ngữ cảnh sâu hơn trong văn bản và đặc biệt phù hợp với ngôn ngữ tự nhiên có sắc thái biểu cảm đa dạng như đánh giá của sinh viên. Kỹ thuật này kiểm tra mức độ tương thích cảm xúc một-nhiều giữa văn bản đầu vào với không gian biểu diễn cảm xúc đã học được qua hàng triệu câu từ. Nhờ đó, hệ thống đạt được độ chính xác cao hơn hẳn so với các hệ thống truyền thống, đặc biệt trong các trường hợp văn bản dài, phức tạp, hoặc chứa các phép ẩn dụ, mỉa mai thường gặp trong phản hồi sinh viên.

1.3. Các ứng dụng hiện nay

Bài toán phân tích cảm xúc từ văn bản là một trong những hướng nghiên cứu quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo. Với khả năng khai thác cảm xúc ẩn chứa trong ngôn ngữ, bài toán này mở ra nhiều tiềm năng ứng dụng trong thực tiễn và đã thu hút sự quan tâm nghiên cứu sâu rộng trong thời gian dài. Các ứng dụng thực tế của hệ thống phân tích cảm xúc có thể kể đến như:

- Hệ thống hỗ trợ chăm sóc khách hàng: tự động phân tích phản hồi văn bản của khách hàng trên các nền tảng như email, khảo sát hoặc mạng xã hội để xác định cảm xúc tiêu cực hoặc tích cực. Qua đó, doanh nghiệp có thể ưu tiên xử lý các phản hồi không hài lòng một cách nhanh chóng và hiệu quả hơn.
- Giám sát dư luận xã hội: trong lĩnh vực truyền thông và quản trị công, hệ thống phân tích cảm xúc có thể giúp theo dõi phản ứng của công chúng trước các chính sách, sự kiện hoặc sản phẩm qua các bài viết, bình luận trên mạng xã hội hoặc diễn đàn.
- Phân tích cảm xúc trong đánh giá sản phẩm hoặc dịch vụ: hỗ trợ tự động phân loại hàng triệu đánh giá trên các sàn thương mại điện tử (như Shopee, Tiki, Amazon) để cung cấp cho người dùng thông tin tóm tắt nhanh về mức độ hài lòng của cộng đồng.
- Hệ thống gợi ý và cá nhân hóa nội dung: trong các nền tảng học trực tuyến hoặc đọc sách, phân tích cảm xúc từ nội dung đánh giá và phản hồi giúp điều chỉnh đề xuất bài học, sách hoặc video phù hợp hơn với trạng thái và sở thích người dùng.

Nhờ những tiềm năng ứng dụng phong phú và thiết thực như vậy, bài toán phân tích cảm xúc từ văn bản không chỉ là một chủ đề nghiên cứu học thuật, mà còn là một giải pháp công nghệ mang tính chiến lược trong thời đại số hiện nay.

1.4. Những khó khăn của bài toán phân tích cảm xúc từ văn bản

Mặc dù phân tích cảm xúc từ văn bản là một bài toán giàu tiềm năng ứng dụng, song quá trình triển khai và nghiên cứu lại gặp phải nhiều khó khăn mang tính kỹ thuật và ngôn ngữ.

Trước hết, ngôn ngữ tự nhiên vốn dĩ phức tạp, đa nghĩa và giàu sắc thái biểu cảm, khiến việc nhận diện cảm xúc không thể chỉ dựa trên từ khóa đơn lẻ. Một câu văn có thể chứa những từ ngữ tích cực nhưng lại biểu thị cảm xúc tiêu cực thông qua cấu trúc mỉa mai, nghịch lý hay chơi chữ. Ví dụ, câu "Môn học thật thú vị – theo cách tệ nhất có thể" thoạt nghe có vẻ tích cực nhưng thực chất mang cảm xúc tiêu cực sâu sắc. Đây là thách thức lớn đối với các mô hình không có khả năng xử lý ngữ cảnh và

các hiện tượng ngôn ngữ phức tạp.

Tiếp theo, độ mơ hồ trong việc gán nhãn cảm xúc cũng là một rào cản đáng kể. Việc đánh giá một văn bản là tích cực, tiêu cực hay trung tính không phải lúc nào cũng rõ ràng, ngay cả đối với con người. Cảm xúc trong ngôn ngữ thường nằm trên phổ liên tục thay vì các nhãn rời rạc, và có thể biến đổi theo ngữ cảnh, chủ đề hoặc thậm chí là phong cách diễn đạt cá nhân. Điều này dẫn đến sự thiếu nhất quán trong dữ liệu huấn luyện, ảnh hưởng đến chất lượng học của các mô hình máy học.

Thứ ba, sự đa dạng về lĩnh vực, đối tượng và văn phong khiến mô hình phân tích cảm xúc cần được điều chỉnh hoặc huấn luyện lại khi chuyển sang miền ứng dụng mới. Chẳng hạn, cảm xúc trong đánh giá phim điện ảnh sẽ khác biệt rõ rệt so với trong phản hồi sinh viên hoặc nhận xét sản phẩm công nghệ. Mô hình học sâu nếu không có dữ liệu đủ lớn và đại diện sẽ dễ gặp phải hiện tượng sai lệch miền (domain shift), làm giảm hiệu quả suy luận.

Ngoài ra, vấn đề xử lý văn bản ngắn và thiếu thông tin ngữ cảnh, điển hình như các đánh giá một dòng, cũng gây khó khăn cho hệ thống. Các văn bản này thường không cung cấp đủ thông tin để mô hình hiểu rõ cảm xúc thực sự của người viết, nhất là khi không có dữ liệu hỗ trợ như lịch sử phản hồi hay thông tin người dùng.

Cuối cùng, chi phí tính toán và yêu cầu về tài nguyên dữ liệu trong các mô hình học sâu hiện đại như BERT, GPT,... cũng là rào cản lớn, đặc biệt đối với các tổ chức giáo dục, doanh nghiệp vừa và nhỏ. Việc xây dựng một hệ thống phân tích cảm xúc hiệu quả không chỉ cần dữ liệu lớn, đa dạng và được gán nhãn chính xác, mà còn đòi hỏi hạ tầng tính toán mạnh để huấn luyện và triển khai mô hình.

Tất cả những yếu tố trên cho thấy rằng, bài toán phân tích cảm xúc từ văn bản, tuy giàu ứng dụng, nhưng đồng thời cũng là một thách thức nghiên cứu mang tính liên ngành, đòi hỏi sự kết hợp giữa ngôn ngữ học, khoa học dữ liệu và học máy để đạt được kết quả chính xác và có tính tổng quát cao.

1.5. Đề xuất hướng tiếp cận trong báo cáo

Trong khuôn khổ nghiên cứu này, đề tài đề xuất ba hướng tiếp cận tổng hợp để giải quyết bài toán phân tích cảm xúc từ văn bản đánh giá của sinh viên, với mục tiêu so sánh hiệu quả giữa các phương pháp tiền xử lý và biểu diễn dữ liệu khác nhau.

Cụ thể, ba hướng tiếp cận chính: (1) TFIDF-Base, (2) FastText-Embed, và (3) PhoBERT-Contextual trong xử lý văn bản đầu vào được triển khai như sau:

Thứ nhất, hướng truyền thống (Traditional) sử dụng kỹ thuật biểu diễn văn bản dựa trên mô hình túi từ (Bag of Words) với cách tiếp cận TF-IDF hoặc TF đơn thuần. Từ đó, xây dựng mô hình đặc trưng dạng đơn vị (unit-level) hoặc vector space for classification (VSFC), nhằm chuyển đổi văn bản thành các vectơ đặc trưng có thể sử dụng trực tiếp trong các mô hình học máy truyền thống. Cách tiếp cận này có ưu điểm đơn giản, dễ triển khai, song thường không nắm bắt được ngữ nghĩa và ngữ cảnh sâu sắc của văn bản.

Thứ hai, hướng biểu diễn ngữ nghĩa bằng từ (Word Embedding) với công cụ FastText – một mô hình học từ vựng do Facebook phát triển, cho phép biểu diễn các từ dưới dạng vectơ liên tục, có khả năng phản ánh tương quan hình thái và ngữ nghĩa. Phương pháp này được kỳ vọng phù hợp với tiếng Việt do FastText có khả năng xử lý từ mới hoặc từ viết sai chính tả nhờ cơ chế chia từ thành các n-gram ký tự.

Thứ ba, hướng tiếp cận biểu diễn ngữ cảnh sâu (Contextual Embedding) sử dụng mô hình tiền huấn luyện PhoBERT – một biến thể của BERT được huấn luyện đặc biệt cho tiếng Việt. PhoBERT cho phép nắm bắt ngữ nghĩa theo ngữ cảnh, nghĩa là một từ trong các câu khác nhau sẽ có biểu diễn khác nhau tùy thuộc vào bối cảnh, từ đó nâng cao độ chính xác trong nhận diện cảm xúc, đặc biệt với các văn bản đánh giá chứa nhiều sắc thái và cấu trúc ngữ pháp phức tạp.

Sau khi hoàn thành quá trình tiền xử lý và trích xuất đặc trưng theo ba hướng nói trên, các biểu diễn dữ liệu sẽ được sử dụng làm đầu vào cho bốn mô hình học máy khác nhau gồm: Random Forest, Support Vector Machine, Extreme Gradient Boosting và Neural Network.

Các mô hình này sẽ được huấn luyện và đánh giá trên cùng tập dữ liệu để so sánh hiệu quả phân loại cảm xúc giữa các kỹ thuật biểu diễn và thuật toán học máy.

Việc sử dụng đồng thời nhiều mô hình với các dạng đặc trưng khác nhau cho phép đưa ra đánh giá toàn diện về độ hiệu quả, tính tổng quát và độ phù hợp của từng hướng tiếp cận đối với bài toán phân tích cảm xúc trong bối cảnh dữ liệu tiếng Việt đặc thù như phản hồi sinh viên.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Khái quát về phân tích cảm xúc – Sentiment Analysis

Phân tích cảm xúc (Sentiment Analysis), còn được biết đến với tên gọi khai thác ý kiến (Opinion Mining), là một nhánh của xử lý ngôn ngữ tự nhiên NLP chuyên nghiên cứu về việc nhận diện, trích xuất và phân loại cảm xúc được thể hiện trong văn bản.

Mục tiêu của phân tích cảm xúc là xác định quan điểm hoặc thái độ mà người viết thể hiện đối với một đối tượng, chủ đề hoặc sự kiện cụ thể trong văn bản, từ đó phục vụ cho nhiều ứng dụng thực tế như phân tích thị trường, đánh giá sản phẩm, đo lường sự hài lòng của khách hàng, và đặc biệt trong bối cảnh giáo dục – phân tích phản hồi sinh viên.[2]

Một trong những cách tiếp cận phổ biến nhất trong phân tích cảm xúc là phân loại văn bản đầu vào theo các nhãn cảm xúc rời rạc. Trong phạm vi nghiên cứu này, hệ thống phân loại cảm xúc được triển khai theo mô hình ba lớp gồm: tích cực (positive), tiêu cực (negative) và trung tính (neutral).

Bảng 2.1. Ví dụ văn bản cho từng loại cảm xúc

Cảm xúc	**Ví dụ văn bản đánh giá của sinh viên**
Tích cực (Positive)	Thầy cô giảng bài dễ hiểu, nhiệt tình hỗ trợ sinh viên.
Trung tính (Neutral)	Buổi học kéo dài 90 phút và kết thúc bằng bài kiểm tra.
Tiêu cực (Negative)	Bài giảng thiếu minh họa, khó tiếp thu kiến thức

Cảm xúc tích cực (Positive): Là nhóm văn bản biểu hiện sự hài lòng, đồng tình hoặc đánh giá cao đối tượng được đề cập. Trong phản hồi sinh viên, các biểu hiện cảm xúc tích cực thường gắn liền với từ ngữ như “hữu ích”, “giáo viên tận tình”, “bài giảng dễ hiểu”, v.v. Các mẫu phản hồi này phản ánh sự thỏa mãn của người học đối với môi trường học tập hoặc chất lượng giảng dạy.

Cảm xúc tiêu cực (Negative): Là các văn bản mang tính chỉ trích, phản đối, bức xúc hoặc không hài lòng. Những văn bản này có thể chứa các từ ngữ như “quá tải”, “thiếu rõ ràng”, “không hiệu quả”,... Trong ngữ cảnh giáo dục, các phản hồi tiêu cực

là tín hiệu quan trọng giúp nhà quản lý cải thiện nội dung giảng dạy, phương pháp tổ chức lớp học hoặc cung cấp hỗ trợ kịp thời cho sinh viên.

Cảm xúc trung tính (Neutral): Là những văn bản không biểu hiện rõ ràng thái độ tích cực hay tiêu cực, thường mang tính mô tả thông tin hoặc đánh giá một cách trung lập. Ví dụ như: "Môn học kéo dài 10 tuần, kết thúc bằng một bài kiểm tra", hoặc "Số lượng sinh viên tham gia khá đông". Lớp trung tính thường khó phân loại do không chứa cảm xúc rõ ràng và có thể dễ gây nhầm lẫn nếu hệ thống không xử lý tốt các dấu hiệu ngữ nghĩa.

Phân loại cảm xúc thành ba lớp giúp hệ thống có khả năng nhận diện đa dạng hơn các sắc thái cảm xúc so với phân loại nhị phân (positive/negative), đồng thời giữ được các đặc điểm văn bản trung lập – vốn chiếm tỷ lệ đáng kể trong các bộ dữ liệu thực tế như đánh giá sinh viên.

Tuy nhiên, mô hình ba lớp cũng đặt ra thách thức cao hơn về mặt phân biệt ranh giới giữa các lớp, đặc biệt giữa lớp trung tính và hai lớp còn lại, đòi hỏi các kỹ thuật trích xuất đặc trưng và mô hình học máy có khả năng nhận diện sắc thái tinh tế trong ngôn ngữ tự nhiên.

2.2. Đặc trưng của văn bản đánh giá trong môi trường giáo dục

Trong bối cảnh giáo dục đại học hiện đại, việc thu thập và khai thác ý kiến phản hồi từ sinh viên đóng vai trò ngày càng quan trọng trong việc cải tiến chất lượng giảng dạy, hoàn thiện chương trình đào tạo và nâng cao trải nghiệm học tập. Các phản hồi này thường được thu nhận dưới dạng văn bản tự do thông qua phiếu khảo sát, cổng thông tin học vụ, hoặc các nền tảng học trực tuyến. Do đó, việc phân tích nội dung các phản hồi sinh viên trở thành một bài toán thiết thực, đòi hỏi phải hiểu sâu về đặc trưng ngôn ngữ và cấu trúc của loại văn bản đặc thù này.

2.2.1. Tính ngắn gọn và phi cấu trúc

Phản hồi của sinh viên thường mang tính ngắn gọn, súc tích và được viết dưới dạng văn bản phi cấu trúc. Nhiều sinh viên có xu hướng sử dụng câu đơn, cụm từ, thậm chí là các biểu thức cảm thán để bày tỏ cảm xúc, thay vì trình bày một cách đầy đủ cú pháp. Ví dụ: "Tốt", "Không hiểu gì hết", "Ổn", "Cô giảng nhanh quá",... Điều

này khiến cho việc xử lý và trích xuất thông tin gặp nhiều khó khăn, đặc biệt trong các kỹ thuật dựa trên cú pháp hoặc ngữ pháp hình thức.

2.2.2. Độ đa dạng trong cách thể hiện cảm xúc

Cùng một nội dung cảm xúc, sinh viên có thể biểu đạt theo nhiều cách khác nhau tùy vào vốn từ, phong cách viết và thái độ cá nhân. Một phản hồi tiêu cực có thể được diễn đạt trực tiếp ("Bài giảng quá khó hiểu") hoặc gián tiếp, thông qua cách dùng từ mỉa mai hoặc ám chỉ ("Học xong vẫn chưa biết học gì"). Độ đa dạng này làm gia tăng tính phức tạp trong việc gán nhãn cảm xúc chính xác, đặc biệt là khi hệ thống phân tích cảm xúc không được huấn luyện trên các ngữ liệu giáo dục đặc thù.

2.2.3. Sự pha trộn giữa cảm xúc và mô tả thông tin

Không ít phản hồi của sinh viên đồng thời chứa cả phần mô tả khách quan và biểu hiện cảm xúc chủ quan. Ví dụ: "Môn học có nhiều bài tập, nhưng giảng viên rất nhiệt tình và hướng dẫn kỹ". Việc bóc tách các thành phần mang cảm xúc và các phần trung lập là một thách thức trong khâu tiền xử lý và phân loại. Hơn nữa, sự đồng hiện của các cảm xúc đối lập trong cùng một câu cũng có thể gây nhiễu cho mô hình nếu không có cơ chế xử lý ngữ cảnh hiệu quả.

2.2.4. Xu hướng sử dụng ngôn ngữ không chuẩn và ký hiệu phi từ vựng

Phản hồi sinh viên trong môi trường số thường chứa các yếu tố ngôn ngữ phi chuẩn như từ viết tắt, teencode, sai chính tả, biểu tượng cảm xúc (emoji), dấu câu không chuẩn, v.v. Ví dụ: "gv dạy ok nhưng ko giải thik rõ", ":((mệt mỏi vì deadline". Những yếu tố này gây khó khăn cho các hệ thống xử lý ngôn ngữ tự nhiên truyền thống, vốn yêu cầu dữ liệu đầu vào chuẩn hóa và nhất quán.

2.2.5. Sự chênh lệch về độ dài và tính không đồng đều của ngữ liệu và tác động của tiếng Việt

Không giống như các tập văn bản có cấu trúc chuẩn như bài báo, tin tức hay bình luận sản phẩm, tập dữ liệu đánh giá sinh viên thường có độ dài không đồng đều. Một số phản hồi chỉ bao gồm một từ duy nhất, trong khi một số khác lại là đoạn văn

dài gồm nhiều câu. Điều này ảnh hưởng trực tiếp đến khả năng trích xuất đặc trưng và yêu cầu mô hình học phải linh hoạt để xử lý tốt cả các mẫu cực ngắn và cực dài.

Cuối cùng, bản chất ngôn ngữ tiếng Việt với đặc trưng ngữ pháp không đánh dấu (non-inflectional), giàu thanh điệu và nhiều từ đồng âm khiến cho việc phân tích ngữ nghĩa và cảm xúc càng trở nên phức tạp. Thêm vào đó, sự thiếu hụt tài nguyên ngôn ngữ học như từ điển cảm xúc tiếng Việt, tập dữ liệu lớn có gán nhãn chuẩn,… cũng là một rào cản đối với việc xây dựng các hệ thống phân tích cảm xúc hiệu quả.

Với các đặc điểm nêu trên, có thể thấy rằng văn bản phản hồi sinh viên trong giáo dục là một dạng ngữ liệu đặc thù, đặt ra nhiều thách thức cho cả giai đoạn tiền xử lý lẫn lựa chọn mô hình phân tích. Do đó, cần có các hướng tiếp cận phù hợp về mặt biểu diễn dữ liệu và xây dựng mô hình học để đảm bảo tính chính xác và khả năng khái quát trong hệ thống phân tích cảm xúc.

2.3. Các phương pháp biểu diễn văn bản

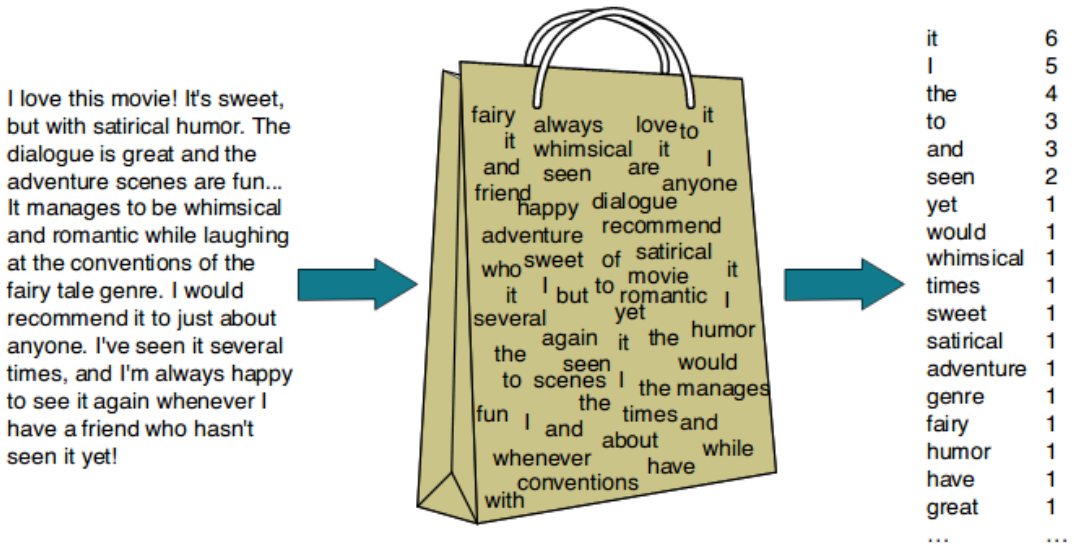
Biểu diễn văn bản (text representation) là một bước trung gian quan trọng trong quá trình xử lý ngôn ngữ tự nhiên (NLP), đặc biệt trong các bài toán học máy như phân loại cảm xúc.

Mục tiêu của bước này là chuyển đổi dữ liệu văn bản thô (dạng chuỗi ký tự) sang dạng số (vector) để các mô hình học máy có thể xử lý được. Chất lượng của biểu diễn văn bản ảnh hưởng trực tiếp đến hiệu năng của mô hình học.

Các phương pháp biểu diễn văn bản có thể được chia thành ba nhóm chính: (1) phương pháp thống kê truyền thống, (2) phương pháp biểu diễn theo không gian ngữ nghĩa (word embedding), và (3) biểu diễn theo ngữ cảnh bằng mô hình ngôn ngữ sâu (contextualized embedding).

2.3.1. Phương pháp truyền thống: Bag-of-Words, TF-IDF

Phương pháp truyền thống như Bag-of-Words (BoW) và TF-IDF biểu diễn văn bản như một tập hợp các từ không có thứ tự. Mỗi tài liệu được ánh xạ thành một vector trong không gian từ vựng như hình 2.2:



Hình 2.2. Ý tưởng Bag of Words

2.3.1.1. Bag-of-Words

Bag of Words là một trong những phương pháp biểu diễn văn bản đơn giản và phổ biến nhất trong xử lý ngôn ngữ tự nhiên.

Đây là kỹ thuật mang tính thống kê, giúp chuyển đổi văn bản sang dạng vector đặc trưng phù hợp với các mô hình học máy truyền thống như SVM, Naive Bayes, hoặc Random Forest. BoW phù hợp với các bài toán phân loại văn bản đơn giản tuy nhiên dễ gây nổ chiều (high-dimensional feature space) nếu tập từ vựng quá lớn và không phân biệt từ đồng nghĩa hay đa nghĩa.

Ý tưởng cốt lõi của BoW là coi một văn bản như một "túi" chứa các từ không quan tâm đến thứ tự xuất hiện mà chỉ tập trung vào tần suất của từ đó trong văn bản.

Đặc điểm của BoW:

- Không quan tâm đến ngữ cảnh hay vị trí của từ trong câu.
- Mỗi từ trong tập từ vựng (vocabulary) được xem như một đặc trưng (feature).

- Mỗi văn bản được biểu diễn thành một vector nhị phân (có/không), đếm (count) hoặc theo tần suất (frequency).

BoW thường đi kèm với việc tiền xử lý dữ liệu: chuyển chữ thường, loại bỏ dấu câu, từ dừng (stop-words), và chuẩn hóa từ.

Document D1	*The child makes the dog happy* the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	*The dog makes the child happy* the: 2, child: 1, makes: 1, dog: 1, happy: 1

↓

	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	**[1,1,1,1,2]**
D2	1	1	1	1	2	**[1,1,1,1,2]**

Hình 2.3. Minh họa cách hoạt động Bag of Word

2.3.1.2. TF-IDF

TF-IDF: cải tiến BoW bằng cách giảm trọng số của các từ phổ biến trong toàn bộ tập văn bản, làm nổi bật các từ mang tính đặc trưng cho mỗi văn bản.

TF(Term frequency) : Tần suất xuất hiện của 1 từ trong 1 document.

TF(t, d) = (Số lần xuất hiện từ t) / (Tổng số từ)

IDF(Invert Document Frequency) : Dùng để đánh giá mức độ quan trọng của 1 từ trong văn bản. Khi tính tf múc độ quan trọng của các từ là như nhau. Tuy nhiên trong văn bản thường xuất hiện nhiều từ không quan trọng xuất hiện với tần suất cao:

- Từ nối : và, hoặc, ...
- Giới từ: ở, trong, của, để, ...
- Từ chỉ định: ấy, đó, nhỉ

Chính vì thế ta cần giảm đi mức độ quan trọng của những từ đó bằng IDF.

IDF(t, D) = log_e(Số văn bản trong tập D /Số văn bản chứa từ t trong tập D)

TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)

Những từ có tf-idf là những từ xuất hiện nhiều trong 1 văn bản này và xuất hiện ít trong văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao trong văn bản (keyword). [4]

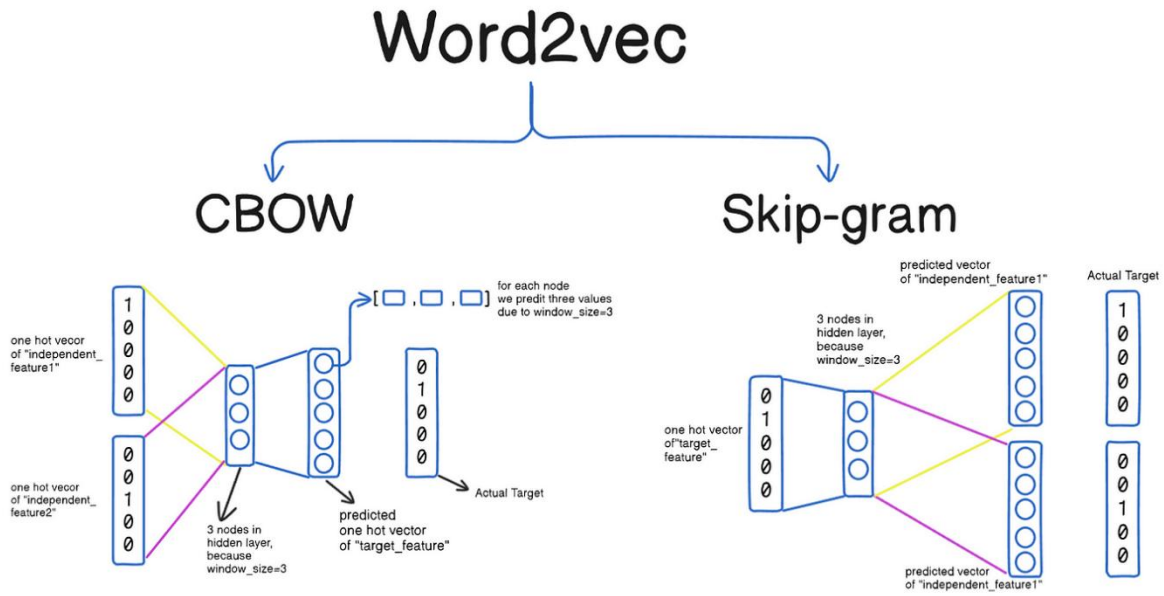
Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Hình 2.4. Minh họa cách hoạt động TF-IDF

2.3.2. Biểu diễn thông qua vector (Word Embedding)

Word embedding là kỹ thuật biểu diễn từ trong không gian vector có chiều thấp, nơi các từ có ý nghĩa tương đồng sẽ có vector gần nhau hơn. Không giống như BoW hay TF-IDF, embedding có khả năng học được các mối quan hệ ngữ nghĩa giữa các từ. Embedding giúp giảm chiều vector và tăng tính khái quát hóa cho mô hình học máy.

Mô hình Word2Vec, được giới thiệu bởi Mikolov và cộng sự tại Google, là một trong những phương pháp phổ biến nhất để học biểu diễn từ (word embeddings).



Hình 2.5. Mô hình Word2Vec

Word2Vec có hai kiến trúc chính: Continuous Bag of Words (CBOW) và Skip-gram.[6]

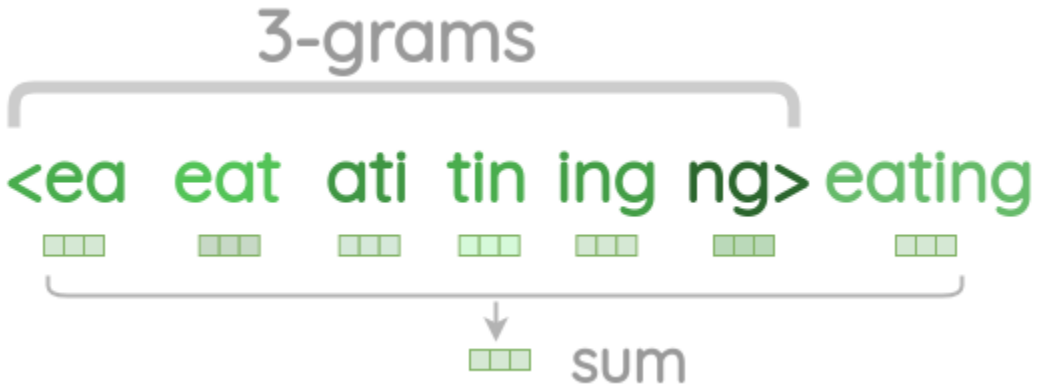
- CBOW là kiến trúc trong đó mô hình học cách dự đoán từ trung tâm (target word) dựa trên các từ xung quanh (context words). Cụ thể, đầu vào là các từ trong cửa sổ ngữ cảnh được biểu diễn dưới dạng one-hot vector, sau đó trung bình qua lớp ẩn để tạo ra vector ẩn. Vector này được sử dụng để dự đoán xác suất của từ trung tâm trong tập từ vựng.

- Skip-gram hoạt động theo hướng ngược lại: dự đoán các từ ngữ cảnh từ một từ trung tâm. Trong mô hình này, đầu vào là từ trung tâm (dưới dạng one-hot vector), sau khi qua lớp ẩn, mô hình sẽ học cách dự đoán các từ ngữ cảnh lân cận. Skip-gram tỏ ra hiệu quả hơn trong việc học đại diện cho các từ hiếm và thường đạt chất lượng cao hơn trong các bài toán ngữ nghĩa tinh vi, mặc dù thời gian huấn luyện có thể dài hơn so với CBOW.

Có thể xem xét câu "The cat sits on the mat".

Với kích thước cửa sổ ngữ cảnh là 3: CBOW sử dụng các từ “The” và “sits” để dự đoán từ trung tâm “cat”. Ngược lại, Skip-gram sẽ sử dụng “cat” để dự đoán các từ xung quanh là “The” và “sits”.

FastText là phiên bản mở rộng của Word2Vec do Facebook phát triển. Nó mở rộng khái niệm nhúng từ truyền thống bằng cách biểu diễn từ dưới dạng các n-gram ký tự, cho phép xử lý hiệu quả các từ ngoài từ vựng (out-of-vocabulary) và các biến thể hình thái. Cách tiếp cận độc đáo này khiến FastText đặc biệt phù hợp với những ngôn ngữ có hình thái phức tạp như tiếng Việt. Mô hình này học cách tạo ra các vector biểu diễn dày đặc (dense vectors) cho từ, các thành phần con (subwords), và n-gram ký tự, giúp nắm bắt các đặc điểm ngữ nghĩa và cú pháp của ngôn ngữ. Việc triển khai FastText rất hiệu quả, cho phép huấn luyện nhanh trên các tập dữ liệu lớn, và vì thế nó được sử dụng phổ biến trong các nhiệm vụ như phân loại văn bản, phân tích cảm xúc và dịch máy. [6]



Hình 2.5. Minh họa cách hoạt động FastText

Dựa trên hình 2.5, FastText biểu diễn mỗi từ như một tập hợp các chuỗi con liên tiếp cố định độ dài (n-grams), ví dụ:

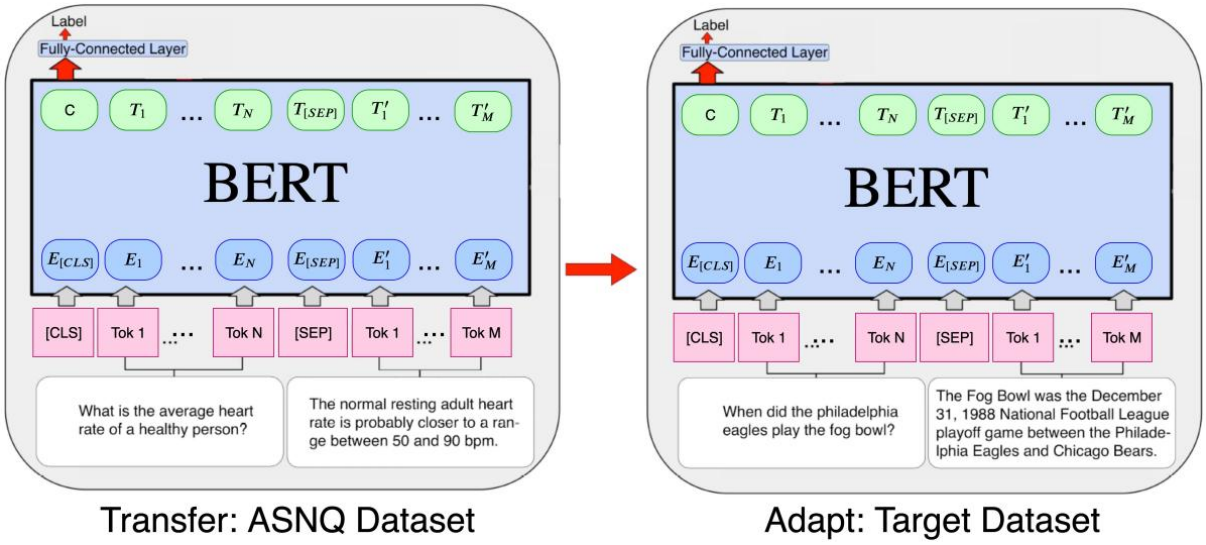
Với 3-gram, từ "eating" sẽ được phân tách thành các đoạn như <ea, eat, ati, tin, ing, ng>. Các ký tự đặc biệt < và > được thêm vào để đánh dấu vị trí đầu và cuối từ, giúp mô hình phân biệt vị trí xuất hiện của các đoạn n-gram. Mỗi n-gram được ánh xạ thành một vector trong không gian đặc trưng, sau đó các vector này được tổng hợp (thường bằng phép cộng) để tạo thành vector biểu diễn từ gốc.

Phương pháp này không chỉ tăng tính khái quát hóa mà còn cải thiện hiệu quả khi làm việc với các ngôn ngữ phức tạp, từ biến thể, lỗi chính tả hoặc từ mới. Nhờ đó, FastText được kỳ vọng phù hợp với các bài toán xử lý ngôn ngữ tự nhiên ở các ngôn ngữ có hình thái phong phú như tiếng Việt.

2.3.3. Biểu diễn theo ngữ cảnh (Contextualized Embedding)

Các phương pháp embedding truyền thống như Word2Vec hay FastText biểu diễn mỗi từ bằng một vector cố định, không phụ thuộc ngữ cảnh. Tuy nhiên, trong ngôn ngữ tự nhiên, cùng một từ có thể mang nhiều nghĩa khác nhau tùy thuộc vào ngữ cảnh. Để khắc phục điểm yếu này, các mô hình ngôn ngữ sâu (deep language models) như BERT được phát triển.

BERT là mô hình tiền huấn luyện sử dụng kiến trúc Transformer, học được các biểu diễn ngữ nghĩa theo ngữ cảnh của từ dựa trên toàn bộ câu văn. Mỗi từ được ánh xạ tới một vector khác nhau tùy thuộc vào ngữ cảnh mà nó xuất hiện.[7]



Hình 2.6. Cách hoạt động mô hình biểu diễn theo ngữ cảnh BERT

Hình 2.6 minh họa trên mô tả quy trình học chuyển giao (transfer learning) bằng BERT, trong đó mô hình được huấn luyện trước trên một tập dữ liệu lớn (ví dụ: ASNQ – tập câu hỏi và câu trả lời trong lĩnh vực sức khỏe) và sau đó được tinh chỉnh (fine-tune) trên một tập dữ liệu đích khác (ví dụ: dữ liệu về trận đấu Fog Bowl trong lịch sử bóng bầu dục Mỹ).

Trong cả hai giai đoạn, đầu vào của BERT bao gồm:

- Câu hỏi và đoạn văn được mã hóa dưới dạng các token (Tok) và được phân tách bởi token đặc biệt [SEP].
- Token [CLS] được đưa vào đầu để tổng hợp thông tin toàn bộ chuỗi và phục vụ cho việc phân loại.

- Mỗi token được biểu diễn bằng các embedding (E) bao gồm thông tin về từ, vị trí và kiểu đoạn (segment).
- Các embedding sau đó được đưa qua các lớp Transformer của BERT để trích xuất đặc trưng ngữ nghĩa (T).
- Cuối cùng, đầu ra từ token [CLS] được đưa vào một lớp fully-connected để dự đoán nhãn (label) như câu trả lời đúng hay mức độ liên quan.

Cơ chế này cho phép BERT học được biểu diễn ngữ nghĩa sâu sắc từ tập dữ liệu nguồn, sau đó điều chỉnh để phù hợp với nhiệm vụ mới mà không cần huấn luyện lại toàn bộ mô hình từ đầu. Nhờ đó, BERT không chỉ tiết kiệm chi phí tính toán mà còn đạt hiệu quả cao trong nhiều bài toán như trả lời câu hỏi, phân loại văn bản, và nhận diện thực thể.

PhoBERT được phát triển bởi nhóm nghiên cứu tại VinAI Research, thuộc Tập đoàn Vingroup (Việt Nam). Đây là mô hình ngôn ngữ tiền huấn luyện đầu tiên cho tiếng Việt dựa trên kiến trúc RoBERTa – một biến thể cải tiến của BERT[8]. PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT. Đây là một trong những mô hình tốt nhất hiện nay trong việc biểu diễn câu tiếng Việt với độ chính xác cao.

Không giống như phương pháp tách từ dựa trên khoảng trắng truyền thống, PhoBERT xử lý đầu vào thông qua cơ chế phân tách từ thành các đơn vị subword, giúp mô hình hoạt động hiệu quả hơn với đặc điểm giàu hình thái và ngôn ngữ phức tạp của tiếng Việt.

Chẳng hạn, câu “Giảng viên rất nhiệt tình và thân thiện.:

- Câu sẽ được tách thành các token như “▁Giảng”, “viên”, “▁rất”, “▁nhiệt”, “tình”, v.v.
- Các token này sau đó được ánh xạ vào không gian vector thông qua các vector nhúng có sẵn, phản ánh thông tin hình thái và ngữ nghĩa ban đầu.
- Tiếp theo, dãy vector đại diện cho câu đầu vào sẽ đi qua nhiều lớp encoder của mô hình Transformer với cơ chế attention hai chiều, cho phép PhoBERT khai thác đồng thời thông tin ngữ cảnh từ cả trái sang phải và

phải sang trái. Điều này đặc biệt quan trọng trong bài toán phân tích cảm xúc, khi cảm xúc không chỉ phụ thuộc vào một từ riêng lẻ mà còn vào cách từ đó tương tác với các từ xung quanh. Ví dụ, cụm "nhiệt tình và thân thiện" được mô hình hiểu là biểu hiện cảm xúc tích cực thông qua sự tương tác giữa các subword và từ bổ nghĩa đi kèm.

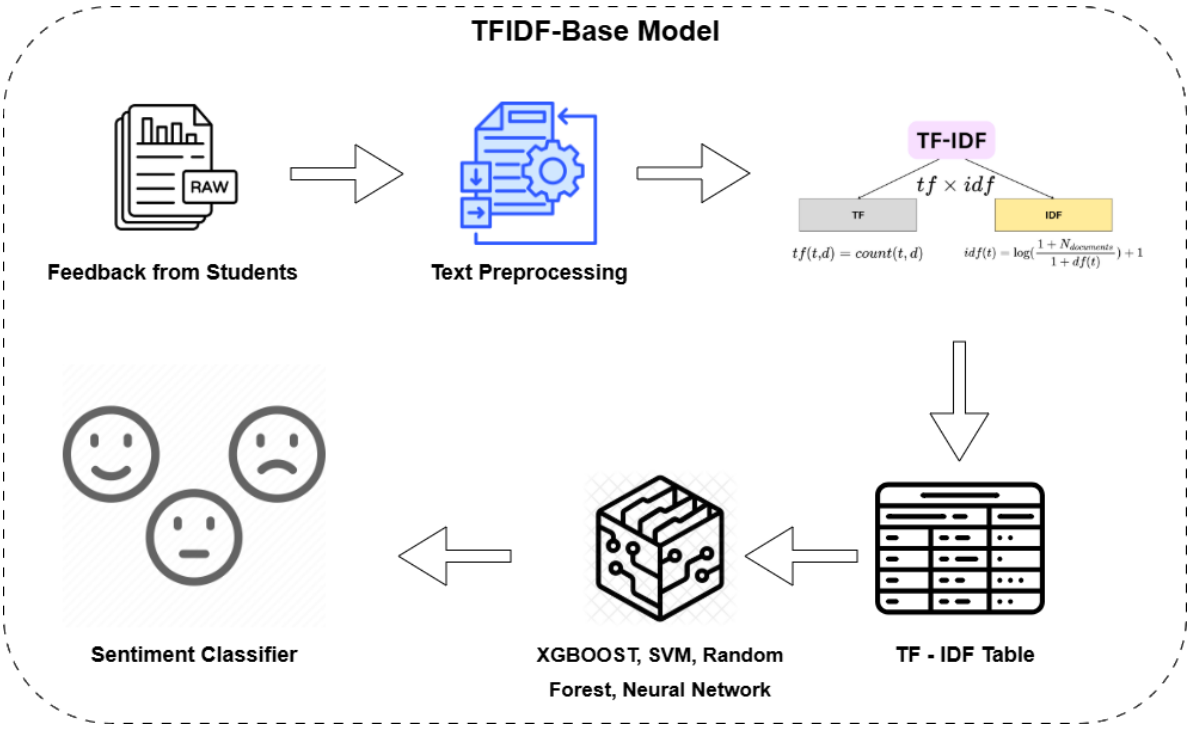
- Cuối cùng, vector đặc trưng toàn câu (đại diện bởi token đặc biệt [CLS]) sẽ được đưa qua tầng phân loại (classification head) để dự đoán nhãn cảm xúc tương ứng, chẳng hạn như tích cực (positive), trung lập (neutral) hoặc tiêu cực (negative).

Nhờ khả năng học biểu diễn ngữ nghĩa giàu ngữ cảnh và khai thác mối quan hệ giữa các thành phần trong câu, PhoBERT đã chứng minh hiệu quả vượt trội trong nhiều tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt, đặc biệt là trong bài toán phân tích cảm xúc từ văn bản đánh giá sinh viên, nơi mà sắc thái cảm xúc thường được biểu đạt một cách đa dạng và không trực tiếp.

CHƯƠNG 3: MÔ HÌNH ĐỀ XUẤT

3.1. Mô hình TFIDF-Base

Trong bối cảnh phân tích cảm xúc dựa trên văn bản đánh giá của sinh viên, một trong những hướng tiếp cận truyền thống phổ biến là sử dụng mô hình biểu diễn đặc trưng bằng TF-IDF. Mô hình này không dựa vào ngữ cảnh như các mô hình embedding hiện đại, nhưng lại mang tính chất trực quan và có khả năng phản ánh mức độ quan trọng của các từ trong tài liệu. Quy trình tổng thể của mô hình được thể hiện trong Hình 3.1.



Hình 3.1. Minh họa mô hình đề xuất 1: TFIDF-Base Model

Ban đầu, các phản hồi văn bản thô từ sinh viên sẽ được đưa vào giai đoạn tiền xử lý bao gồm loại bỏ ký tự đặc biệt, chuyển đổi chữ thường, tách từ và loại bỏ stopwords.

Sau khi tiền xử lý, văn bản được biến đổi thành vector đặc trưng bằng cách áp dụng phương pháp TF-IDF, trong đó mỗi từ được gán trọng số dựa trên tần suất xuất hiện trong văn bản so với toàn bộ tập dữ liệu. Cụ thể, giá trị TF phản ánh tần suất của một từ trong văn bản, trong khi IDF đánh giá mức độ hiếm gặp của từ trong toàn bộ tập tài liệu. Việc nhân hai giá trị này tạo ra một đại diện vector hóa có khả

năng làm nổi bật các từ mang tính đặc trưng cao cho mỗi văn bản.

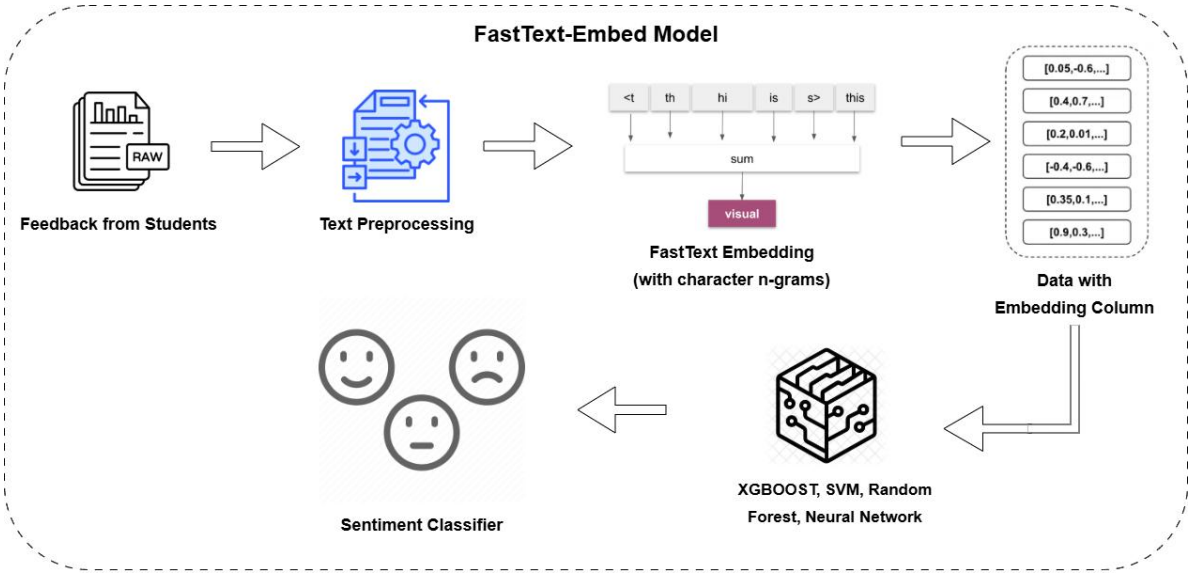
Cuối cùng, dữ liệu sẽ được đưa vào các mô hình học máy truyền thống như SVM , Random Forest, Neural Network hoặc XGBoost nhằm thực hiện phân loại cảm xúc thành ba lớp: tích cực (positive), trung tính (neutral) và tiêu cực (negative). Mặc dù phương pháp này không tận dụng được thông tin ngữ cảnh hoặc quan hệ giữa các từ, nhưng nó cung cấp một đường cơ sở đáng tin cậy để so sánh với các mô hình embedding hiện đại hơn như FastText hoặc PhoBERT trong cùng một bài toán.

3.2. Mô hình FastText-Embed

Trong những năm gần đây, các mô hình nhúng từ (word embedding) đã chứng minh hiệu quả vượt trội so với các phương pháp truyền thống trong các bài toán xử lý ngôn ngữ tự nhiên (NLP), đặc biệt là trong ngữ cảnh tiếng Việt – một ngôn ngữ có cấu trúc giàu hình thái.

Mô hình FastText, được phát triển bởi Facebook AI Research, là một trong những phương pháp embedding tiêu biểu khai thác thông tin ngữ âm và cấu trúc từ bằng cách sử dụng n-gram ký tự. Trong nghiên cứu này, nhóm đề xuất mô hình FastText-Embed Model nhằm chuyển đổi các văn bản đánh giá của sinh viên thành các vector đặc trưng giàu ngữ nghĩa để phục vụ bài toán phân loại cảm xúc.

Quy trình hoạt động của mô hình được trình bày trong Hình 3.2.



Hình 3.2. Minh họa mô hình đề xuất 2: FastText-Embed Model

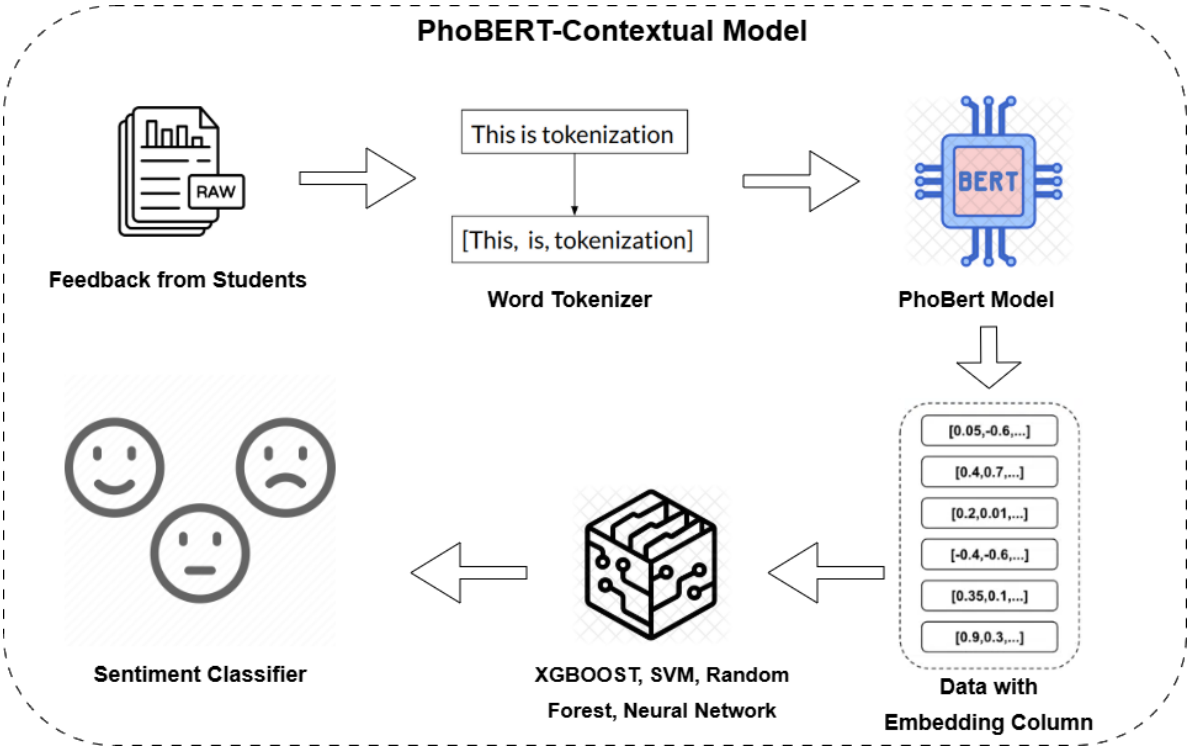
Dữ liệu đầu vào là các phản hồi văn bản thô của sinh viên sẽ trải qua giai đoạn tiền xử lý bao gồm chuẩn hóa văn bản, xóa ký tự đặc biệt, loại bỏ stopwords và tách từ. Sau bước tiền xử lý, văn bản được đưa vào mô hình FastText để ánh xạ mỗi từ thành một vector liên tục trong không gian thực.

Điểm đặc biệt của FastText so với Word2Vec là nó không chỉ biểu diễn mỗi từ như một đơn vị riêng lẻ, mà còn phân tích từ thành các n-gram ký tự con, ví dụ như từ "this" sẽ được phân rã thành <th, thi, his, is>... và tổng hợp lại thành vector của từ. Nhờ đó, mô hình có khả năng xử lý tốt các từ mới (OOV – out-of-vocabulary), vốn là một thách thức lớn trong tiếng Việt.

Các vector embedding thu được từ FastText sẽ được lưu trữ kèm theo nhãn cảm xúc tương ứng, tạo thành bộ dữ liệu học. Sau đó, dữ liệu embedding sẽ được sử dụng để huấn luyện các mô hình phân loại như XGBoost, SVM, Random Forest hoặc mạng nơ-ron. Mục tiêu cuối cùng của mô hình là gán nhãn cảm xúc cho mỗi phản hồi sinh viên với ba mức: tích cực, trung tính hoặc tiêu cực. Mô hình FastText-Embed nhờ tận dụng đặc trưng n-gram ký tự không chỉ giúp tăng độ chính xác phân loại mà còn mang lại tính linh hoạt và khả năng khái quát tốt hơn cho hệ thống so với phương pháp TF-IDF.

3.3. Mô hình PhoBERT-Contextual

Trong bối cảnh phát triển mạnh mẽ của các mô hình ngôn ngữ theo hướng biểu diễn ngữ cảnh (contextualized representation), PhoBERT đã nổi lên như một giải pháp tối ưu cho tiếng Việt, được huấn luyện dựa trên kiến trúc Transformer của BERT và dữ liệu tiếng Việt quy mô lớn. Trong nghiên cứu này, nhóm xây dựng mô hình PhoBERT-Contextual nhằm khai thác triệt để khả năng biểu diễn ngữ nghĩa ngữ cảnh của PhoBERT để phục vụ cho bài toán phân loại cảm xúc từ phản hồi của sinh viên. Các biểu diễn embedding từ PhoBERT sau đó được kết hợp với các thuật toán học máy truyền thống như SVM, Random Forest, XGBoost và Neural Network để thực hiện phân loại. Quy trình hoạt động chi tiết của mô hình được minh họa trong Hình 3.3.



Hình 3.3. Minh họa mô hình đề xuất 3:Mô hình PhoBert-Contextual

Mô hình PhoBERT-Contextual được thiết kế nhằm tận dụng sức mạnh của biểu diễn ngữ cảnh trong bài toán phân loại cảm xúc văn bản tiếng Việt. Quy trình xử lý của mô hình gồm các bước như sau:

Đầu vào của hệ thống là các đoạn phản hồi (feedback) từ sinh viên, được thu

thập ở dạng văn bản thô (raw text). Các văn bản này trước tiên được đưa vào bộ token hóa (word tokenizer), có nhiệm vụ chia tách văn bản thành các đơn vị từ (tokens) phù hợp với định dạng đầu vào của mô hình PhoBERT.

Sau khi token hóa, dữ liệu văn bản được đưa vào mô hình PhoBERT, một mô hình ngôn ngữ tiền huấn luyện theo kiến trúc BERT, được tối ưu riêng cho tiếng Việt. PhoBERT sinh ra các vector nhúng ngữ cảnh (contextualized embeddings) cho từng văn bản, biểu diễn đặc trưng ngữ nghĩa của câu dựa trên ngữ cảnh toàn cục.

Các vector embedding thu được có dạng các chuỗi số thực (như [0.05, 0.6, …]) và được gom thành cột đặc trưng (embedding column). Tập dữ liệu với cột embedding này được sử dụng làm đầu vào cho các mô hình học máy như: XGBoost, SVM, Random Forest hoặc Neural Network.

Cuối cùng, mô hình học máy thực hiện phân loại cảm xúc cho từng văn bản phản hồi thành một trong ba nhãn: tích cực (positive), tiêu cực (negative), hoặc trung tính (neutral). Nhờ việc kết hợp biểu diễn ngữ cảnh mạnh mẽ từ PhoBERT với các thuật toán phân loại hiệu quả, mô hình PhoBERT-Contextual hứa hẹn sẽ cho thấy hiệu suất vượt trội trong bài toán phân tích cảm xúc tiếng Việt.

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. Dữ liệu thực nghiệm.

Trong nghiên cứu này, nhóm sử dụng bộ dữ liệu "Synthetic Vietnamese Students Feedback Corpus" được cung cấp trên nền tảng Kaggle.[9]

Bộ dữ liệu bao gồm hơn 10.000 mẫu phản hồi giả lập của sinh viên Việt Nam, được thiết kế nhằm hỗ trợ các nghiên cứu về phân tích cảm xúc và phân loại chủ đề trong văn bản tiếng Việt.

Mỗi mẫu dữ liệu bao gồm ba trường thông tin: "sentence" (câu phản hồi), "sentiment" (nhãn cảm xúc với ba mức độ: tích cực, tiêu cực, trung tính) và "topic" (chủ đề liên quan như học tập, giảng viên, cơ sở vật chất, v.v.).

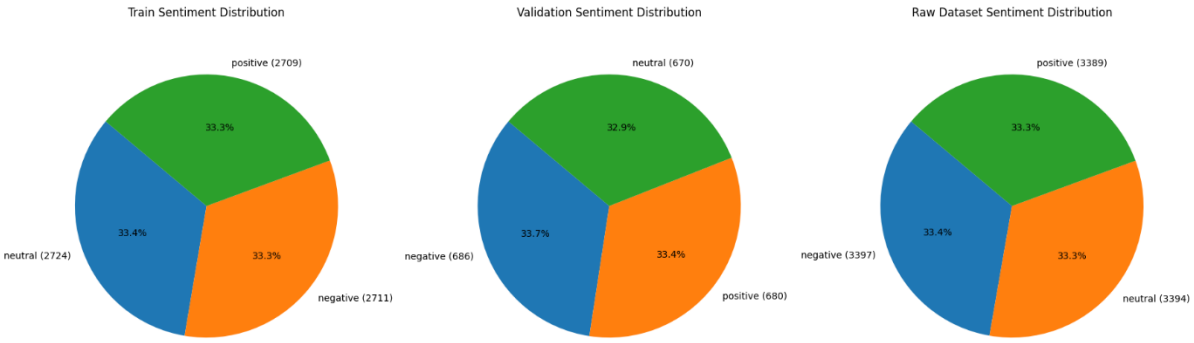
Bảng 4.1. Ví dụ các mẫu trong dataset

No.	Sentence	Sentiment	Topic
1	Đội ngũ bảo trì quá thưa thớt dẫn đến không đảm bảo được chất lượng sửa chữa thiết bị.	Negative	Facility
2	Giảng viên này quá yêu cầu với sinh viên khiến họ không thể tự do học tập.	Negative	Lecturer
3	Các phòng thí nghiệm tại trường được trang bị đầy đủ và là nơi lý tưởng để học tập và nghiên cứu.	Positive	Facility
4	Thầy luôn đặt mục tiêu rõ ràng cho học viên nên học viên biết mình phải làm gì để đạt được điểm cao.	Positive	Lecturer
5	Chương trình học của trường rất đa dạng và bao gồm nhiều ngành học hấp dẫn.	Neutral	Curriculum
6	Có những sinh viên đánh giá khách quan và đưa ra các ý kiến xây dựng sau mỗi bài học.	Neutral	Others

Để phục vụ quá trình huấn luyện và đánh giá mô hình, bộ dữ liệu được chia thành hai tập: tập huấn luyện (80%) và tập kiểm tra (20%). Việc sử dụng bộ dữ liệu này giúp đảm bảo tính nhất quán trong đánh giá hiệu quả của các mô hình học máy trong nhiệm vụ phân tích cảm xúc và phân loại chủ đề trên văn bản tiếng Việt.

4.2. Phân tích khám phá dữ liệu

4.2.1. Phân bố nhãn



Hình 4.2. Pie chart tỷ lệ Sentiment trong tập Dataset, train và validation

Bảng 4.3. Số lượng cụ thể từng biến Sentiment trong tập Dataset, train và validation

No.	File	Sentiment	Quantity
1	Dataset	Neutral	3397
		Negative	3394
		Positive	3389
2	Train	Neutral	2724
		Negative	2711
		Positive	2709
3	Validation	Neutral	686
		Negative	680
		Positive	670

Dựa trên thống kê hình 4.2 và bảng 4.3, tập dữ liệu được sử dụng trong bài toán đánh giá cảm xúc từ văn bản đánh giá của sinh viên bao gồm ba nhãn cảm xúc chính: positive (tích cực), negative (tiêu cực) và neutral (trung tính).

Tổng cộng, tập dữ liệu chứa 10.680 mẫu, trong đó số lượng mẫu theo từng nhãn gần như được phân bố đều: neutral chiếm 3.397 mẫu (≈33.3%), negative chiếm 3.394 mẫu (≈33.4%) và positive là 3.389 mẫu (≈33.3%). Điều này cho thấy dữ liệu được cân bằng tốt giữa các loại cảm xúc, hạn chế nguy cơ mô hình bị thiên lệch trong quá trình huấn luyện.

Tập huấn luyện (train) bao gồm 8.144 mẫu, với tỉ lệ các nhãn neutral (2.724 mẫu, ≈33.4%), negative (2.711 mẫu, ≈33.3%) và positive (2.709 mẫu, ≈33.3%) được phân bố rất đồng đều.

Tập kiểm tra (validation) gồm 2.036 mẫu, cũng duy trì sự cân bằng giữa các nhãn: neutral (686 mẫu, ≈33.7%), negative (680 mẫu, ≈33.4%) và positive (670 mẫu, ≈32.9%).

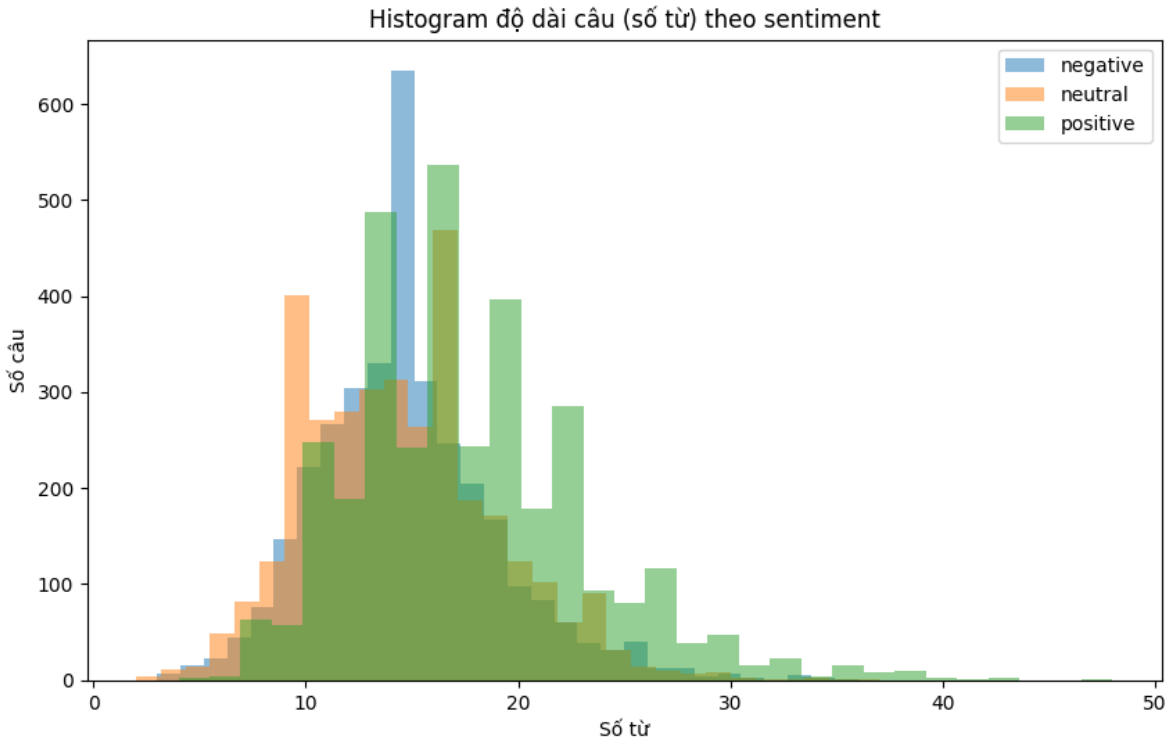
Như vậy, cả ba tập dữ liệu (toàn bộ, train và validation) đều được phân chia hợp lý, hỗ trợ quá trình huấn luyện và đánh giá mô hình một cách khách quan và hiệu quả.

4.2.2. Độ dài câu theo từng sentiment

Một trong những đặc điểm quan trọng cần được xem xét trong quá trình tiền xử lý và phân tích khám phá dữ liệu là độ dài của câu, đặc biệt trong các bài toán phân loại cảm xúc văn bản.

Việc khảo sát độ dài giúp hiểu rõ hơn về cấu trúc ngôn ngữ của tập dữ liệu, đồng thời cung cấp thông tin hữu ích cho việc lựa chọn mô hình học sâu, đặc biệt là khi có giới hạn độ dài đầu vào (input length) như trong các mô hình transformer.

Hình 4.4 minh họa phân bố độ dài câu (tính theo số lượng từ) tương ứng với ba nhãn cảm xúc: tiêu cực (negative), trung tính (neutral) và tích cực (positive), thông qua biểu đồ histogram chồng lớp.



Hình 4.4. Histogram độ dài câu theo sentiment

Về tổng thể, phần lớn các câu trong cả ba nhóm sentiment đều có độ dài dao động trong khoảng từ 10 đến 25 từ, cho thấy dữ liệu phản ánh khá tốt đặc điểm của các câu phản hồi ngắn gọn trong môi trường giáo dục. Các câu tích cực (positive) có xu hướng phân bố rộng hơn về phía độ dài lớn, tức là thường có số từ cao hơn so với hai nhóm còn lại. Điều này có thể lý giải bởi người học thường diễn đạt chi tiết hơn khi phản hồi tích cực.

Ngược lại, các câu trung tính (neutral) thường ngắn hơn rõ rệt, tập trung chủ yếu trong khoảng 8–15 từ, cho thấy đặc điểm trung lập, đơn giản, mang tính mô tả nhiều hơn là đánh giá. Trong khi đó, các câu tiêu cực (negative) có xu hướng tập trung cao ở khoảng 13–17 từ, với mức độ phân tán thấp hơn nhóm tích cực, phản ánh việc người học thường nêu ý kiến tiêu cực một cách trực tiếp, súc tích.

Nhìn chung, phân bố độ dài câu theo từng loại cảm xúc cho thấy sự khác biệt đáng kể, gợi ý rằng đặc trưng hình thức (như số từ) có thể mang một phần thông tin hữu ích trong việc hỗ trợ mô hình phân loại cảm xúc.

4.3. Giới thiệu môi trường và quy trình thực nghiệm

Trong nghiên cứu này, mọi thực nghiệm sẽ được thực hiện trên thư viện Scikit-Learn dựa trên ngôn ngữ Python. Môi trường thực nghiệm được triển khai trên các máy tính có cấu hình Intel Core i9-13500H 2.6 GHz và hệ điều hành Windows 11.

Các thí nghiệm được thiết kế nhằm đánh giá hiệu suất của các mô hình khác nhau trong việc phân tích cảm xúc dựa trên đánh giá của sinh viên như trên bảng 4.5. Cụ thể, bảy nhóm thực nghiệm đã được triển khai để kiểm tra khả năng xử lý dữ liệu văn bản cũng như hiệu suất phân loại cảm xúc dựa trên các phương pháp học máy và phương pháp xử lý văn bản.

Bảng 4.5. Quy trình thực nghiệm

No	Model	Input	Output	Mục đích
1	TFIDF-Base	Train Validation	Sentiment classification into: positive, neutral, or negative.	Biến văn bản thành đặc trưng số thông qua TF-IDF và huấn luyện mô hình học máy cơ bản để phân loại sentiment.
2	FastText-Embed			Biểu diễn câu bằng cách trung bình các vector được huấn luyện trước từ FastText, sau đó dùng mô hình học máy để phân loại.
3	PhoBERT-Contextual			Biểu diễn ngữ nghĩa sâu của câu tiếng Việt bằng mô hình ngữ cảnh hóa PhoBERT, rồi dùng vector này cho bài toán phân loại sentiment.

4.4. Độ đo đánh giá hiệu suất mô hình

Trong báo cáo này, các độ đo được sử dụng để đánh giá hiệu quả của mô hình phân loại bao gồm: precision (độ chính xác theo dự đoán), recall (độ bao phủ), F1-score, support, và accuracy (độ chính xác tổng thể). Mỗi độ đo phản ánh một khía cạnh khác nhau trong hiệu năng của mô hình:

- **Precision** (Độ chính xác theo dự đoán): là tỷ lệ giữa số lượng dự đoán đúng thuộc một lớp cụ thể (True Positive) so với tổng số mẫu mà mô hình

dự đoán thuộc lớp đó (gồm cả đúng và sai). Precision đánh giá mức độ "tin cậy" khi mô hình gán nhãn cho một lớp.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Trong đó:

- TP (True Positive – Dương tính đúng): Là số lượng mẫu thực sự thuộc lớp dương (ví dụ: lớp "Positive") và mô hình cũng dự đoán đúng là lớp đó.
- FP (False Positive – Dương tính sai): Là số lượng mẫu thực sự không thuộc lớp dương nhưng mô hình dự đoán nhầm là lớp đó. Ví dụ: Một phản hồi trung tính nhưng mô hình lại đoán là tích cực.

→ Trong số các dự đoán là dương, bao nhiêu là đúng thực sự.

- **Recall** (Độ bao phủ): là tỷ lệ giữa số lượng dự đoán đúng thuộc một lớp (True Positive) so với tổng số mẫu thực sự thuộc lớp đó (bao gồm cả những mẫu bị bỏ sót). Recall phản ánh khả năng "bao quát" của mô hình đối với từng lớp.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

Trong đó:

- FN (False Negative – Âm tính sai): Là số lượng mẫu thực sự thuộc lớp dương nhưng mô hình lại đoán sai là lớp khác. Ví dụ: Một phản hồi tích cực nhưng mô hình đoán là trung tính hoặc tiêu cực.

→ Trong số các mẫu thực sự là dương, mô hình đã phát hiện được bao nhiêu.

- **F1-score**: là trung bình điều hòa giữa precision và recall, nhằm cân bằng hai chỉ số trên. F1-score đặc biệt hữu ích trong các bài toán phân loại có sự mất cân bằng về số lượng mẫu giữa các lớp.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

  - Nếu F1-score cao, tức là mô hình không chỉ dự đoán đúng nhiều (precision cao), mà còn không bỏ sót nhiều nhãn thật (recall cao).
  - Nếu precision cao mà recall thấp (hoặc ngược lại), thì F1-score sẽ kéo về mức trung bình → giúp ta đánh giá toàn diện hơn.
- Support: là số lượng mẫu thực tế thuộc mỗi lớp trong tập kiểm tra. Cho biết mức độ đóng góp của mỗi lớp vào chỉ số tổng thể. Nếu một lớp có support rất nhỏ, thì dù F1-score cao cũng chưa chắc có ý nghĩa lớn toàn cục.
  - Ví dụ: Nếu support của lớp "Positive" là 680, nghĩa là có 680 phản hồi thực sự là tích cực trong tập test.
- Accuracy: Là tỷ lệ giữa tổng số mẫu được mô hình dự đoán đúng và tổng số mẫu trong toàn bộ tập dữ liệu kiểm tra.

$$\text{Accuracy} = \frac{\text{Tổng số dự đoán đúng}}{\text{Tổng số mẫu}}$$

4.5. Thực nghiệm 1: Mô hình TFIDF-Base

Bảng 4.6. Kết quả thực nghiệm mô hình TFIDF-Base

Model		Predicted Label			Metrics				
		Negative	**Neutral**	**Positive**	**Precison**	**Recall**	**F1-Score**	**Support**	**Accuracy**
RF	**Negative**	661	16	9	0.91	**0.96**	0.94	686	0.82
	Neutral	47	472	151	0.77	0.7	0.73	670	
	Positive	19	128	533	**0.77**	**0.78**	0.78	680	
SVM	**Negative**	660	20	6	0.94	**0.96**	**0.95**	686	0.81
	Neutral	28	499	143	0.73	0.74	0.74	670	
	Positive	12	168	500	**0.77**	0.74	0.75	680	
XGB	**Negative**	659	23	4	0.94	**0.96**	**0.95**	686	**0.84**
	Neutral	31	514	125	**0.76**	**0.77**	**0.77**	670	
	Positive	13	136	531	0.8	**0.78**	**0.79**	680	
NN	**Negative**	635	40	11	**0.95**	0.93	0.94	686	0.8
	Neutral	20	499	151	0.7	0.74	0.72	670	
	Positive	10	169	501	0.76	0.74	0.75	680	

Bảng 4.6 trình bày kết quả thực nghiệm của bốn mô hình phân loại văn bản sử dụng biểu diễn đặc trưng TFIDF-Base, bao gồm Random Fores, SVM, XGBoost và Neural Network.

Nhìn chung, mô hình XGBoost đạt hiệu suất tổng thể cao nhất với độ chính xác (accuracy) là 0.84, vượt trội so với các mô hình còn lại. Cụ thể, XGBoost thể hiện tốt ở cả ba lớp, đặc biệt là lớp Negative với F1-Score là 0.94 và lớp Positive với F1-Score là 0.79 – cao nhất trong số các mô hình.

Mô hình RF và SVM có hiệu suất tương đối đồng đều, với accuracy lần lượt là 0.82 và 0.81. Tuy nhiên, RF có ưu thế hơn ở lớp Negative (F1 = 0.94) nhưng lại thể hiện yếu ở lớp Neutral (F1 = 0.73), cho thấy mô hình này khó phân biệt các văn bản trung tính với hai cực còn lại. Mô hình SVM cũng gặp vấn đề tương tự ở lớp Neutral, dù có độ chính xác cao ở lớp Negative (F1 = 0.95).

Trong khi đó, mô hình NN cho kết quả thấp nhất với accuracy là 0.80, dù có F1-Score cao ở lớp Negative (0.94) nhưng kém ổn định ở hai lớp còn lại, đặc biệt là lớp Neutral (F1 = 0.73) – tương tự các mô hình khác.

Bảng 4.6 không chỉ phản ánh hiệu quả của các mô hình học máy khác nhau mà còn cho thấy tiềm năng rõ rệt của biểu diễn đặc trưng TF-IDF trong việc xử lý và phân loại dữ liệu văn bản. Mặc dù TF-IDF là một phương pháp truyền thống và không khai thác ngữ cảnh sâu như các mô hình embedding hiện đại (ví dụ: PhoBERT, Word2Vec), nhưng kết quả thực nghiệm cho thấy nó vẫn đủ mạnh để hỗ trợ các mô hình học máy cổ điển đạt độ chính xác và F1-Score tương đối cao.

Cụ thể, với chỉ đặc trưng TF-IDF, các mô hình như XGBoost và Random Forest đã đạt accuracy lần lượt là 0.84 và 0.82, đồng thời thể hiện khả năng phân biệt hiệu quả giữa ba lớp cảm xúc (Negative, Neutral, Positive). Điều này chứng tỏ rằng TF-IDF có khả năng khai thác được các từ khóa mang tính đặc trưng cao trong văn bản, từ đó giúp mô hình học máy xây dựng được ranh giới phân lớp rõ ràng.

Kết luận, TF-IDF vẫn là một phương pháp biểu diễn văn bản đơn giản nhưng hiệu quả trong các bài toán phân loại văn bản, đặc biệt khi kết hợp với các mô hình học máy mạnh như XGBoost. Với ưu điểm về tốc độ xử lý và khả năng mở rộng, TF-IDF là lựa chọn phù hợp cho các bài toán phân loại văn bản quy mô lớn hoặc trong môi trường yêu cầu tính toán hiệu quả.

4.6. Thực nghiệm 2: Mô hình FastText-Embed

Bảng 4.7. Kết quả thực nghiệm mô hình FastText-Embed

Model		Predicted Label			Metrics				
		Negative	**Neutral**	**Positive**	**Precison**	**Recall**	**F1-Score**	**Support**	**Accuracy**
RF	**Negative**	610	41	35	0.84	0.89	0.86	686	0.75
	Neutral	69	418	183	0.71	0.62	0.67	670	
	Positive	49	126	505	0.7	0.74	0.72	680	
SVM	**Negative**	602	46	38	0.89	0.88	0.88	686	0.77
	Neutral	54	427	189	0.71	0.64	0.67	670	
	Positive	20	129	531	0.7	0.78	0.74	680	
XGB	**Negative**	619	35	32	0.9	0.9	0.9	686	0.78
	Neutral	46	474	150	0.71	0.71	0.71	670	
	Positive	22	154	504	0.73	0.74	0.74	680	
NN	**Negative**	625	36	25	**0.93**	**0.91**	**0.92**	686	**0.81**
	Neutral	32	490	148	**0.74**	**0.73**	**0.73**	670	
	Positive	14	138	528	**0.75**	**0.78**	**0.76**	680	

Dựa trên Bảng 4.7 – kết quả thực nghiệm với mô hình FastText-Embed, ta nhận thấy rằng mặc dù FastText là một kỹ thuật embedding hiện đại do Facebook phát triển, có khả năng biểu diễn từ ngữ bằng vector ngữ nghĩa và xử lý được từ chưa xuất hiện (OOV) thông qua subword, nhưng hiệu quả tổng thể của mô hình khi áp dụng cho văn bản tiếng Việt lại không vượt trội so với phương pháp TF-IDF truyền thống.

Cụ thể, mô hình Neural Network với FastText đạt độ chính xác cao nhất là 0.81, trong khi mô hình XGBoost chỉ đạt 0.78 và Random Forest thấp hơn nữa (0.75). Ngoài ra, điểm F1-Score trung bình của các lớp với FastText cũng cho thấy sự dao động lớn, đặc biệt là lớp Neutral luôn bị phân loại kém chính xác nhất (F1 thấp nhất từ 0.67 đến 0.74), tương tự hiện tượng xảy ra với TF-IDF, nhưng mức độ phân biệt vẫn kém hơn.

So sánh giữa hai bảng kết quả (Bảng 4.6 và Bảng 4.7), ta thấy rõ rằng mô hình

sử dụng đặc trưng TF-IDF nhìn chung cho kết quả ổn định và cao hơn so với mô hình sử dụng FastText embedding.

Cụ thể, mô hình XGBoost với TF-IDF đạt độ chính xác cao nhất (0.84), trong khi với FastText chỉ đạt 0.78. Tương tự, các chỉ số F1-Score trên từng lớp trong mô hình TF-IDF cũng có xu hướng cao và đồng đều hơn, đặc biệt là ở lớp Negative và Positive. Trong khi đó, các mô hình sử dụng FastText thường gặp khó khăn trong việc phân loại lớp Neutral, dẫn đến F1-Score thấp và kéo giảm độ chính xác chung.

Kết luận, dù FastText có ưu thế về mặt ngữ nghĩa và đã chứng minh hiệu quả trên nhiều ngôn ngữ lớn, tuy nhiên trong bối cảnh xử lý tiếng Việt – một ngôn ngữ có đặc thù về cấu trúc từ ghép, dấu thanh, và sự đa dạng biểu đạt – thì mô hình FastText-Embed vẫn chưa thể hiện được ưu thế rõ rệt so với TF-IDF. Điều này gợi ý rằng việc sử dụng embedding hiện đại cần được điều chỉnh hoặc huấn luyện lại trên dữ liệu tiếng Việt chuyên biệt, hoặc kết hợp với mô hình học sâu phù hợp hơn để phát huy toàn bộ tiềm năng – điều nhóm tiến hành trong thực nghiệm 3.

4.7. Thực nghiệm 3: Mô hình PhoBert-Contextual

Bảng 4.8. Kết quả thực nghiệm mô hình PhoBert-Contextual

Model		**Predicted Label**			**Metrics**				
		Negative	**Neutral**	**Positive**	**Precison**	**Recall**	**F1-Score**	**Support**	**Accuracy**
RF	**Negative**	668	10	8	0.91	0.97	0.94	686	0.81
	Neutral	54	460	156	0.75	0.69	0.72	670	
	Positive	16	145	519	0.76	0.76	0.76	680	
SVM	**Negative**	666	19	1	**0.96**	0.97	**0.97**	686	**0.84**
	Neutral	18	496	156	0.77	0.74	0.75	670	
	Positive	7	131	542	0.78	0.8	**0.79**	680	
XGB	**Negative**	665	14	7	0.94	0.97	0.96	686	**0.84**
	Neutral	30	505	135	0.77	**0.75**	**0.76**	670	
	Positive	9	134	537	**0.79**	0.79	**0.79**	680	
NN	**Negative**	682	3	1	0.94	**0.99**	0.96	686	**0.84**
	Neutral	34	471	165	**0.8**	0.7	0.75	670	
	Positive	12	116	552	0.77	**0.81**	**0.79**	680	

Bảng 4.8 trình bày kết quả thực nghiệm của mô hình PhoBERT-Contextual, một mô hình ngôn ngữ hiện đại được huấn luyện chuyên biệt cho tiếng Việt, khi kết hợp với các mô hình học máy khác nhau. Có thể thấy rằng, PhoBERT cho kết quả vượt trội hơn hẳn so với các mô hình sử dụng đặc trưng TF-IDF (thực nghiệm 1) và FastText (thực nghiệm 2) trước đó.

Cụ thể, tất cả các mô hình kết hợp với PhoBERT đều đạt độ chính xác (accuracy) rất cao, từ 0.81 đến 0.84, trong đó các mô hình SVM, XGBoost và Neural Network đều đạt accuracy tối đa 0.84 – mức cao nhất trong toàn bộ các thực nghiệm. Đặc biệt, F1-Score trên từng lớp đều cao và ổn định, cho thấy mô hình có khả năng phân loại cân bằng giữa ba lớp Negative, Neutral và Positive. Ví dụ, mô hình SVM đạt F1-Score lên đến 0.97 ở lớp Negative, trong khi vẫn duy trì mức F1 từ 0.75 đến 0.79 cho hai lớp còn lại.

Điểm đáng chú ý là mô hình NN kết hợp PhoBERT không chỉ đạt độ chính xác cao (0.84) mà còn thể hiện F1-Score tốt nhất ở lớp Positive (0.79) và Neutral (0.76), đồng thời duy trì độ ổn định giữa các lớp – điều mà các mô hình TF-IDF hoặc FastText không đạt được.

Kết luận, PhoBERT-Contextual là phương pháp biểu diễn văn bản hiệu quả nhất trong toàn bộ các thực nghiệm, khi được kết hợp với các mô hình học máy mạnh như SVM, XGBoost hoặc Neural Network. Mô hình này không chỉ tận dụng ngữ cảnh ngôn ngữ sâu của tiếng Việt mà còn giúp cải thiện đáng kể hiệu quả phân loại. Nhờ đó, PhoBERT chứng tỏ là lựa chọn ưu việt nhất trong các phương pháp biểu diễn và phân loại văn bản tiếng Việt hiện nay.

4.8. Nhận xét và kết luận thực nghiệm

Trong phần thực nghiệm, ba phương pháp biểu diễn đặc trưng văn bản gồm TF-IDF, FastText embedding và PhoBERT-Contextual đã được đánh giá trên cùng một tập dữ liệu phân loại cảm xúc với ba lớp: Negative, Neutral và Positive. Mỗi phương pháp được kết hợp với bốn mô hình học máy gồm: Random Forest (RF), SVM, XGBoost và Neural Network (NN). Các kết quả được trình bày lần lượt trong các bảng 4.6, 4.7 và 4.8.

- **Mô hình TFIDF-Base**

Kết quả ở Bảng 4.6 cho thấy mô hình sử dụng TF-IDF – dù là phương pháp truyền thống – vẫn đạt hiệu quả cao, đặc biệt khi kết hợp với XGBoost, cho độ chính xác lên đến 0.84. Nhìn chung, các mô hình sử dụng TF-IDF đạt F1-Score ổn định, nhất là ở hai lớp Negative và Positive. Đây là minh chứng cho thấy TF-IDF vẫn là một công cụ mạnh mẽ trong xử lý văn bản tiếng Việt, nơi mà các tín hiệu từ khóa bề mặt đóng vai trò quan trọng.

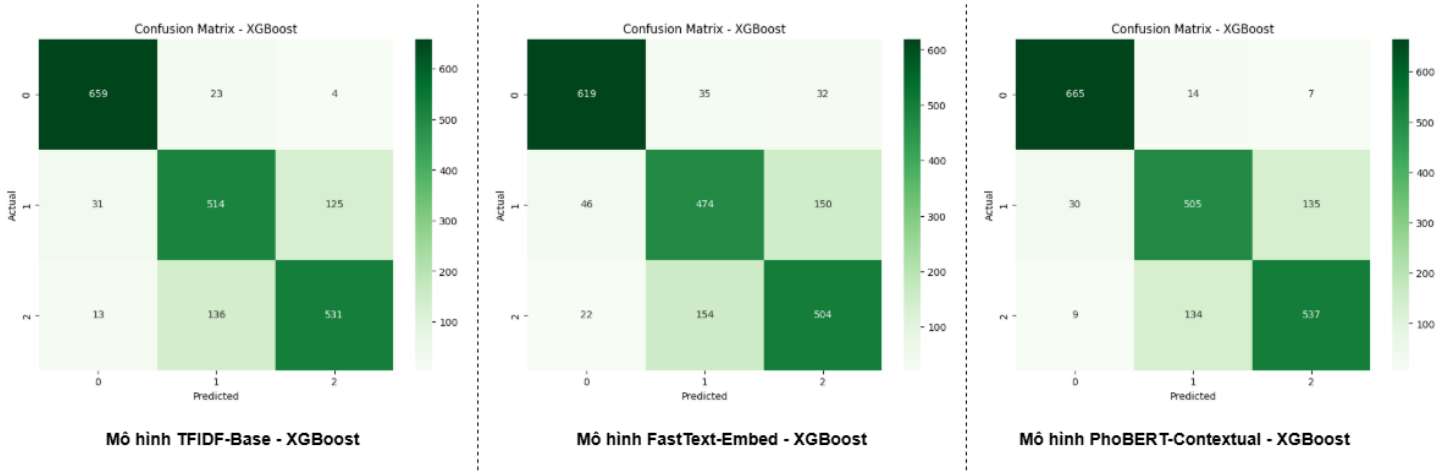
- **Mô hình FastText-Embed**

Ngược lại, kết quả trong Bảng 4.7 cho thấy FastText – mặc dù là kỹ thuật embedding hiện đại do Facebook phát triển – lại chưa phát huy được hiệu quả vượt trội trong ngữ cảnh tiếng Việt. Các mô hình kết hợp FastText chỉ đạt độ chính xác cao nhất là 0.81 (NN), còn lại chủ yếu dao động trong khoảng 0.75–0.78. Đặc biệt, mô hình thường gặp khó khăn trong việc phân loại lớp Neutral, dẫn đến F1-Score thấp ở lớp này. Điều này có thể đến từ việc embedding FastText chưa được tinh chỉnh phù hợp với ngữ cảnh và đặc trưng ngôn ngữ tiếng Việt.

- **Mô hình PhoBERT-Contextual**

Bảng 4.8 cho thấy PhoBERT-Contextual là phương pháp biểu diễn vượt trội nhất trong các thực nghiệm. Khi kết hợp với các mô hình học máy, PhoBERT consistently đạt độ chính xác cao nhất (0.84) ở ba mô hình SVM, XGBoost và NN. Ngoài ra, các chỉ số F1-Score cao và đồng đều trên cả ba lớp cảm xúc cho thấy PhoBERT có khả năng hiểu ngữ cảnh và đặc điểm ngôn ngữ tiếng Việt tốt hơn nhiều so với các phương pháp còn lại. Mô hình NN kết hợp PhoBERT còn cho F1-Score cao nhất ở lớp Positive (0.79), đồng thời duy trì mức cân bằng giữa các lớp.

- **Confusion Matrix**



Hình 4.9. Confusion Matrix giữa các phương pháp xử lý văn bản kết hợp XGBoost

Để đánh giá hiệu quả của các phương pháp biểu diễn văn bản kết hợp với mô hình phân loại XGBoost trong nhiệm vụ phân tích cảm xúc, chúng tôi so sánh ma trận nhầm lẫn thu được từ ba mô hình: TFIDF-Base, FastText-Embed, và PhoBERT-Contextual trên Hình 4.9. Kết quả cho thấy mô hình PhoBERT-Contextual - XGBoost đạt hiệu quả phân loại cao nhất trên cả ba lớp cảm xúc: tiêu cực, trung tính, và tích cực.

Cụ thể, đối với lớp tiêu cực, mô hình PhoBERT dự đoán đúng 665 mẫu, vượt trội so với TFIDF (659) và FastText (619). Với lớp trung tính, PhoBERT đạt 505 mẫu đúng, tương đương với TFIDF (514) và tốt hơn FastText (474). Đối với lớp tích cực, PhoBERT tiếp tục dẫn đầu với 537 mẫu được phân loại đúng, trong khi TFIDF và FastText lần lượt đạt 531 và 504.

Xét về nhầm lẫn, mô hình TFIDF-Base gặp khó khăn trong việc phân biệt giữa cảm xúc trung tính và tích cực, với số lượng nhầm lẫn cao giữa hai lớp (125 mẫu trung tính bị nhầm sang tích cực và 136 mẫu tích cực bị nhầm sang trung tính). FastText-Embed cũng có xu hướng nhầm lẫn giữa các lớp liền kề, đặc biệt là giữa tiêu cực và trung tính, khiến hiệu suất phân loại bị giảm. Ngược lại, PhoBERT-Contextual, nhờ khả năng biểu diễn ngữ nghĩa theo ngữ cảnh, đã làm giảm đáng kể các lỗi nhầm lẫn này.

Tổng thể, mô hình PhoBERT-Contextual khi kết hợp với XGBoost chứng

minh được ưu thế vượt trội so với các phương pháp biểu diễn truyền thống, đặc biệt trong việc xử lý ngữ nghĩa phức tạp trong văn bản tiếng Việt.

- **Kết luận tổng quan**

Tổng hợp các kết quả thực nghiệm, có thể khẳng định rằng:

1. PhoBERT-Contextual là phương pháp hiệu quả nhất trong bài toán phân loại văn bản tiếng Việt, nhờ khả năng hiểu ngữ nghĩa và bối cảnh ngôn ngữ sâu sắc.
2. TF-IDF vẫn chứng minh được tính hiệu quả cao, đặc biệt khi kết hợp với các mô hình học mạnh như XGBoost, đồng thời dễ triển khai và không yêu cầu tài nguyên lớn.
3. FastText tuy hiện đại nhưng chưa thể hiện được ưu thế khi áp dụng vào tiếng Việt, và cần được tinh chỉnh hoặc kết hợp với các chiến lược huấn luyện phù hợp hơn.

Kết quả này là cơ sở quan trọng để lựa chọn kỹ thuật biểu diễn văn bản phù hợp trong các bài toán xử lý ngôn ngữ tự nhiên tiếng Việt sau này.

KẾT LUẬN VÀ ĐỀ XUẤT

Trong nghiên cứu này, nhóm đã tiến hành so sánh hiệu quả của ba phương pháp biểu diễn văn bản kết hợp với các mô hình học máy trong nhiệm vụ phân tích cảm xúc tiếng Việt, bao gồm TFIDF-Base, FastText-Embed và PhoBERT-Contextual.

Kết quả thực nghiệm cho thấy PhoBERT- Contextual là mô hình có hiệu suất vượt trội nhất, với khả năng học biểu diễn ngữ cảnh sâu sắc, góp phần cải thiện rõ rệt độ chính xác phân loại trên cả ba nhãn cảm xúc. Điều này khẳng định tiềm năng của các mô hình ngôn ngữ tiền huấn luyện dựa trên Transformer như PhoBERT trong xử lý văn bản tiếng Việt.

Tuy nhiên, kết quả ma trận nhầm lẫn cũng cho thấy PhoBERT vẫn chưa đạt được mức độ tối ưu tuyệt đối. Đặc biệt, vẫn tồn tại sự nhầm lẫn đáng kể giữa hai nhãn positive (tích cực) và neutral (trung tính). Nguyên nhân có thể đến từ sự mờ ranh giới ngữ nghĩa giữa các phản hồi mang tính tích cực nhẹ và trung lập, vốn là đặc trưng ngôn ngữ thường gặp trong các phản hồi của sinh viên. Bên cạnh đó, việc xử lý văn bản theo câu đơn lẻ có thể làm mất đi ngữ cảnh mở rộng, gây ảnh hưởng đến quá trình phân loại.

Từ những quan sát trên, hướng phát triển tiếp theo của nghiên cứu là:

1. Tích hợp thông tin ngữ cảnh liên đoạn để giảm mơ hồ ngữ nghĩa và tăng độ phân biệt giữa các cảm xúc gần nhau.
2. Kết hợp các kỹ thuật tăng cường dữ liệu (data augmentation) để cải thiện tính đa dạng và khả năng tổng quát của mô hình.
3. Đồng thời, xem xét áp dụng các mô hình ngôn ngữ đa nhiệm (multi-task learning) để đồng thời học cảm xúc và chủ đề, giúp tăng cường thông tin bổ trợ trong phân loại.

Những hướng đề xuất này không chỉ nhằm nâng cao hiệu quả phân tích cảm xúc tiếng Việt, mà còn mở rộng tiềm năng ứng dụng trong các hệ thống phản hồi người dùng, đánh giá chất lượng dịch vụ và hỗ trợ ra quyết định giáo dục.

TÀI LIỆU THAM KHẢO

[1] Chính phủ Việt Nam, “Giáo dục và đào tạo phải lấy học sinh làm trung tâm, nhà trường là nền tảng, giáo viên là động lực,” Báo Chính phủ, 2022. [Online]. Available: https://baochinhphu.vn/gddt-phai-lay-hoc-sinh-lam-trung-tam-nha-truong-la-nen-tang-giao-vien-la-dong-luc-102292224.htm

[2] FPT.AI, “Phân tích cảm xúc (Sentiment Analysis) là gì?”, FPT.AI, [Online]. Available: https://fpt.ai/vi/bai-viet/sentiment-analysis.

[3] AIML.com Research, “Advantages and Disadvantages of Bag of Words (examples, video),” AIML.com. [Online]. Available: https://aiml.com/advantages-disadvantages-bag-of-words/

[4] A. Agrawal, “How to process textual data using TF-IDF in Python,” freeCodeCamp, Oct. 2020. [Online]. Available: https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/

[5] F. Omarzai, “Word2Vec: CBOW & Skip-Gram in depth,” Medium, Jan. 2023. [Online]. Available: https://medium.com/@fraidoonomarzai99/word2vec-cbow-skip-gram-in-depth-88d9cc340a50

[6] A. Mishra, “Getting started with FastText word embeddings,” Amitness, Feb. 2020. [Online]. Available: https://amitness.com/posts/fasttext-embeddings

[7] J. Alammar, “The Illustrated BERT, ELMo, and co.,” jalammar.github.io, 2018. [Online]. Available: https://jalammar.github.io/illustrated-bert/

[8] D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” arXiv preprint arXiv:2003.00744, 2020. Available: https://arxiv.org/abs/2003.00744

[9] T. Leon, “Synthetic Vietnamese Students Feedback Corpus,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/toreleon/synthetic-vietnamese-students-feedback-corpus